



**UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO**

FACULTAD DE CIENCIAS

**De Trinitate Physicae: Una introducción intuitiva al
formalismo de la Mecánica Cuántica**

T E S I S

**QUE PARA OBTENER EL TÍTULO DE:
MATEMÁTICO**

P R E S E N T A:

ALFREDO BANTE OLVERA



**DIRECTOR DE TESIS:
Dr. MICHO ĐURĐEVIC
2019**

Ciudad Universitaria, Ciudad de México. 2019



Universidad Nacional
Autónoma de México



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Foreword

In a world full of misleading presentations of quantum mechanics, relativity, and the physical sciences in general, it is of highest importance to learn things properly and from the right sources. That does not imply, however, that an intuitive, reachable approach to the objects of study is not needed. Quantum mechanics is how we study the physics of the very small, nothing more. It is challenging, for it defies a lifetime of accumulated intuition. It is mesmerising, for it offers a new outlook on physics and science itself, a whole different worldview. It is complex, for it involves a large spectrum of seemingly unconnected branches of knowledge. A lot of learning material is available to popularise both quantum mechanics and the Theory of Relativity, as well as abundant specialised literature; there are not, however, many texts that can offer a “bridge” between a basic state of comprehension and that of a senior undergraduate student, perhaps even a way to link intuitive concepts to certain specialised technicalities. One often finds oneself midway between two intellectual “levels” when trying to understand some aspect of physics. Intuition plays an important role in understanding; it is therefore of uttermost importance to develop a feeling for physics, just as it is to handle confidently the associated mathematical apparatus. Data without interpretation is meager; mathematics without intuition is meaningless. The purpose of this text is to expose the technical formalities of an introductory version of quantum physics in such a way as to link them to a wide, structured perspective of the subjects. It is a starting point to understand and demystify the (perhaps) foundational, conceptual-necessities of such theory.

This text is divided into three chapters; each of them was written with a slightly different purpose. Despite the way they are all interconnected, they can be read separately. The first chapter is intended to explain, with an intuitive approach, the actual experiments that motivated the creation of this new theory. Whether by means of pure reasoning, mathematical insight, or simply experimental data, quantum mechanics brings together a wide range of notions into a whole new perspective, a brand new approach to the study of physical phenomena. Instead of describing each of the experiments separately, and then trying to assemble the different conclusions into a whole, structured theory, a discussion on the notions of waves, particles, and the so-called ‘wave-particle duality’ intertwines these experimental conclusions with the historic standpoints of the different scientists involved. The rest of chapter one is dedicated to the purpose of presenting the most appealing conclusions of quantum theory, and a taste of its mathematical formalism, wherein its beauty lies.

Just as chapter one presents a conceptual introduction to the physics of quantum mechanics, chapter two could be described as an intuitive introduction to its mathematical formalism. Most of modern physics (and its applications) is done with the aid of quantum mechanics, but its mathematical foundations are often hard to grasp since they involve at least a decent level of calculus, linear algebra, group theory, probability theory, and complex analysis. Furthermore, most of it is translated into Dirac’s notation, so the student interested in quantum theory might find it useful to handle it confidently from the very beginning. In

contrast to any other physical description of Nature, the mathematical formalisation of quantum mechanics did not come from empirical knowledge, but vice versa. Most of its notions come from theoretical explorations, and only afterwards were they tested experimentally. This chapter is dedicated to the purpose of explaining, with the aid of examples and visualisation, the mathematics one needs to learn quantum mechanics comfortably and efficiently.

Chapter three is an outline of the main, controversial discussions that prevail, even to this day, on the way quantum theory should be interpreted. From the beginning, it was heavily scrutinised both by physicists, and philosophers; despite its counter-intuitive nature, it continues to predict physical phenomena correctly, and with great precision. In this chapter, the main postures are presented and further discussed in a simple, straightforward way.

In summary, this text condenses a few years of conscientious thinking about the best way to convey the details of such an intricate theory. Throughout the first years of an average undergraduate programme, one usually has introductory courses on several branches of physics like analytical mechanics, thermodynamics, electromagnetism, etc; each of them is a universe in itself, with their own notation, examples, and fields of application. Likewise, one learns several branches of mathematics, and then tries to bring all of these together to understand the language of quantum physics. This endeavour seems sometimes overwhelming, and most texts focus on an in-depth analysis of one particular branch, without trying to interrelate them. The goal of this text is precisely to assemble these various topics and present them in a way that is useful for someone trying to grasp the nature of quantum theory, to ‘connect the dots,’ so to speak. Hopefully, someone with an interest in an intuitive approach to the main subjects of quantum mechanics will find it pleasant.

For the reader interested in a summarised version of the text, the paragraphs marked with the symbol ¶ follow a parallel sequence that this reader might find perhaps more welcoming than the entire text.

Contents

0 Præfatio

1	Introductory Topics	9
§1	History of the Atomic Model	9
§1.1	Atomism	9
§1.2	Dalton	10
§1.3	J.J. Thomson	12
§1.4	Rutherford	13
§1.5	Bohr	14
§1.6	Schrödinger	15
§2	Double Slit Experiment	16
§2.1	The Problem with Light	18
§2.2	What we Understand as Matter	18
§2.3	What we Understand as Waves	20
§2.4	Light as a Wave	22
§2.5	Light as a Particle	24
§3	Some Necessary Historical Notes	27
§3.1	Blackbody Radiation	27
§3.2	What does it Mean to be <i>Discrete</i> ?	30
§4	The Experiment: Electron Diffraction	32
§4.1	The First Quantum Conjecture	32
§4.2	Davisson & Germer	34
§4.3	The Wave Equation	35
§5	Waves and Particles: One, Both, or Neither	39
§6	So, What Does it Look Like?	41
§6.1	Quantum Theory	41
§6.2	The <i>Size</i> of the Wave Function: A Probability Distribution	42
§6.3	Measurements	46
§6.4	The Typical Problem (I)	49
§6.5	Digression: Harmonics of the <i>Mikrokosmos</i>	52
§6.6	The Typical Problem (II): The Solution to the Schrödinger Equation	54
§7	Why is it so Intriguing?	56
§7.1	Superposition	56

	§7.2	Quantisation	57
	§7.3	“Collapse” of the Wave Function	59
	§7.4	Heisenberg’s Inequalities: “Uncertainty”	60
	§7.5	Quantum Tunneling	66
	§7.6	EPR	68
§8		Interpretations	72
	§8.1	The Ensemble Interpretation	73
	§8.2	The Copenhagen Interpretation	73
§9		A Somewhat Satisfying Justification	74
	§9.1	Præliminaris	74
	§9.2	Condon & Cassen: A Primitive Formalism	76
2		Mathematics behind Quantum Theory	78
§1		Vectors & Vector Spaces	83
	§1.1	What <i>is</i> a vector?	83
	§1.2	The Pythagorean Theorem: Norm	84
	§1.3	Matrix Multiplication	85
	§1.4	Transposition	86
	§1.5	What to Imagine when Talking about \mathbb{R}^n	87
	§1.6	More on Matrix Multiplication	88
	§1.7	Inner Product and its Geometrical Meaning	88
§2		Linear Transformations	90
	§2.1	Basis Vectors	91
	§2.2	Linearity	92
	§2.3	To Illustrate...	94
	§2.4	Examples in \mathbb{R}^3	97
	§2.5	The Determinant: A Useful Criterion I	99
	§2.6	More on Vector Transposition	101
	§2.7	The Determinant: A Useful Criterion II	102
	§2.8	Coordinate Transformations	104
	§2.9	A Reformulation of Known Analytical Geometry	108
	§2.10	Complex Numbers	112
	§2.11	Matrices: Some Examples of Different Interpretations	117
	§2.12	Dual Spaces	120
	§2.13	Outer Products and Projectors	121
	§2.14	Function Spaces	122
	§2.15	Taylor Series	124
	§2.16	Linear Combinations and Infinite Basis	126
	§2.17	Inner Product	127
§3		Group Theory	129
	§3.1	Topology of \mathbb{R}^2 and \mathbb{R}^3 : Grasping an Intuition	129
	§3.2	Continuity: A Criterion	131
	§3.3	The Space of Transformations	132
§4		Probability	134
	§4.1	Mean Value: Why is it not Enough?	135
	§4.2	Expectation Values, Variance, and Standard Deviation	138

§4.3	Distributions	140
§5	<i>Die Zusammenfassung</i>	141
§5.1	The Space of Quantum Mechanics	141
§5.2	Expectation Values	142
§5.3	The Schrödinger Equation and Eigenvalue Equations	143
3	The Problem	145
§1	Interpretations	149
§1.1	Collective <i>vs.</i> Individual	150
§1.2	The Problem <i>per se</i>	150
§2	Axiomatisation: The de Broglie Hypothesis	151
§2.1	Mathematics <i>vs.</i> Physics	153
§3	Bohmian Mechanics	154
§4	Further Discussion (Some Final Thoughts)	155
§4.1	<i>Ignoramus et ignorabimus</i>	155
4	Appendices	158
§1	Appendix 1: The Structure of Space and Time	160
§1.1	Flatland	160
§1.2	What <i>is</i> a Shadow?	163
§1.3	<i>Space-Time</i>	165
§1.4	Curvature	168
§1.5	Physics & Geometry	171
§2	Appendix 2: <i>Hydrodynamic Quantum Analogue</i>	176
§2.1	What are HQA's?	176
§2.2	Fluid Mechanics	176
§2.3	Setup for the Experiments	178
§2.4	The Analogy: Connecting the Dots	179
§3	Appendix 3: Arriving at the Schrödinger Equation	183
§3.1	Wave Function	183
§3.2	Finding the Time and Spatial Derivatives	184
§4	Appendix 4: Reminders of Algebraic Definitions	188
§4.1	Structures with one Binary Operation on a Set	188
§4.2	Structures with two Binary Operations on a Set	189
§4.3	Structures with two Binary Operations on two Sets	190
§4.4	Structures with three Binary Operations on two Sets	191
§4.5	In General	192
5	Bibliography	193

Chapter 0

Præfatio: A Brief Discussion on Knowledge

More often than not, one learns science with the aid of mathematics. Despite an obstinate persistence to avoid this by distilling it from equations or numbers, one ends up paraphrasing the underlying mathematical relations in a way that mathematics just seems to be absent, but it never really is. What is not often done, however, is a critical analysis of the way we use, learn and think of mathematics. How is it that the human brain, capable of logical reasoning, *understands* mathematics?

There are essentially three aspects involved in an holistic appreciation of mathematics, the first one being the natural process of logical reasoning, i.e. the gradual construction of concepts through the assumption of a set of pre-established facts. This is the way of *school mathematics*, in which we are taught a series of basic, intrinsically primal concepts like that of *numbers* and the subsequent idea of *counting*; we further use them to build up gradually the well-known operations of addition, multiplication, subtraction, and division.

With arithmetic, and further development, we manage to edify algebra, calculus, and so forth. One could say this is a *forward* way of learning, since we are provided with the elementary, constituent blocks, and we are expected to *build* mathematics with them.

The second one is perhaps more *logically* obvious, however historically unachievable; it does not happen often in history that one unequivocally defines concepts first, with perfect understanding of their meaning and extent, and explore their consequences afterwards. A partial, intuitive understanding of concepts comes first, exploration comes next, and formalisation comes third. One does not question the given foundational entities during the learning process; quite the contrary, they are assumed and gradually worked upon. E.g. a question about the ontological nature of numbers, related to what numbers or operations *are*, is out of the scope of school mathematics, not just the answer to such questions, but the questions per se.

Because of the impossibility of permanently questioning the underlying structure at each and every step of the learning process, one must accept that knowledge is built on seemingly unstable bases. It is only after the whole of the arithmetical and algebraic apparatus has been extensively worked upon that

one can take the time to *step back*, and ponder upon the formal meanings of its constituent elements. If the former part of the learning process is in fact forward, then the latter one could be said to go *backwards*, since it focuses on the task of providing solid foundational backgrounds, making each and every piece perfectly unambiguous, well-defined, and the whole system consistent.

Going *backwards* means asking questions, often very simple ones with increasingly complex answers; it means revising the consistency of the definitions we gave to the primal concepts and the interrelations between them, without necessarily trying to develop the operational apparatus further. The formal consistency of mathematics was pondered upon during the 19th century, mainly as a matter of formality, but clear answers only appeared until the 20th century. Proofs of consistency are a major subject of study within mathematical logic, and their epistemological importance extends much further away than one might guess at first glance¹.

Lastly, it is the predominant role that depiction has had throughout the history of mathematics that makes it worthy of being separately taken into consideration. What we have learnt and discovered with the only help of intuition and drawings is extraordinarily abundant; furthermore, there is profound meaning underlying many mathematical ideas and, despite their complexity, it can be easily conveyed through pictoric depiction. Depiction is associated to geometry as a treatise on Earth and space, at least in its origins and development in ancient Greece, but the true depth of its importance goes much further than that. The mathematical depiction of space led to the possibility of taking different, abstract spaces into consideration. It was only until the 19th century when the first considerations of non-Euclidean geometry were explored.

Euclid wrote a treatise on geometry as a formal description of the space around us, and it successfully portrays the structure of physical space as we perceive it. Any other possibility would have seemed inevitably wrong, since it did not describe space accurately. To our surprise, the discovery of spherical, elliptic, and hyperbolic geometry meant a huge step forward in the investigation of both real physical space and abstract mathematics².

Diagrams and depictions have a heuristic value, but they do not, per se, constitute a formal language. Our formal mathematical language was, however, developed by means of diagrammatic depiction. Take, for instance, the *equal* sign “=,” which was invented by the Welsh mathematician Robert Recorde in 1557 as a means to avoid the “tedious repetition” of the phrase *is equal to*. The mere usage of a formal mathematical language exhibits the importance of synthesising the wholeness of scientific thought; to this extent, equations are, per se, a symbolic and diagrammatic means to convey sense. Despite the formal meaning of their constituent symbols, they end up being used schematically or diagrammatically and can thus, to a certain extent, be considered as formalised

¹For an intuitive approach on the problem of consistency, Gödel’s theorems, and their proofs see: NAGEL, E., et al (2001). *Gödel’s Proof*. New York University Press.

²An excellent introduction to analytical geometry and non-Euclidean geometries can be found in: BRACHO, J. (2009). *Introducción analítica a las geometrías*. Fondo de Cultura Económica

depictions³.

By recreating the process of learning mathematical reasoning, we can reconstruct its development throughout history and vice versa. We learn mathematics almost precisely as humanity discovered (or invented) it; e.g. no one approaches any subject for the first time by being provided with the most precise, formal definitions (or solutions) for the most generalised cases known; we are shown, in a very meticulous and succinct way, the path that led to the basic ideas, their further development, and finally to their resolutions in the form of mathematical syntax. In a way, we relive thousands of years of history during our personal learning process. In the case of elementary mathematics, we are presented with a plethora of notions, none of which we are allowed to question or argue. Instead of pondering about their ontological substance, we are supposed to grow *acquainted* with them. In a way, they remain dogmatic assertions until we have the (mathematical) maturity to examine them closely.

Once we become familiar with the operational methods, once we can fluently handle the functional complexities to a certain degree, we go no further into unpleasant, unnecessary complications. E.g. once we know how to multiply three or four digits numbers, we find it pointless and perhaps annoying to try to multiply five or six digits ones; we opt for the possibility of different, not necessarily more complicated, knowledge⁴.

From the fully built mechanism of basic arithmetic, and with a bit of further analysis and abstraction, we reach a point of deep *conceptual necessity*; we identify an epistemological need to see beyond the numbers and become sensitive enough as to see the hidden *properties* of our arithmetical system. That which lies beneath such familiar operations, is only to be found in the fertile realm of algebra. Via this abstraction we bear witness to some of the deep interrelations and properties of mathematical entities. From the most elementary algebraic notions, and all the way to the most abstract ones (like those pertaining to group theory, linear algebra, Lie Theory, etc), they all provide the mathematical language with great means to express even the most unsuspected connections.

Once these relations are grasped, but now with algebra as a tool, and not as an object of study, the progressive, constructive process evolves. Our “forward” progression to achieve mathematical maturity unfolds as we refine our comprehension of deeper mathematical notions. More and more, powerful methods appear as new mathematical objects come into existence. To unravel the hidden and seemingly mysterious properties of such entities with ever-improving tools and techniques becomes one of the process’ main goals.

³*Feynman diagrams* are a good example of how a diagrammatic depiction can convey formal meaning, and connect an abstract thought with formal knowledge. Not only do they formally represent abstract thought, they do it with a formal, operational logic, i.e. they are a formal language from which one can deduce unequivocal results.

⁴An holistic view of physics involves experimentation, a process I have not included here as an element of a complete appreciation of mathematics. It is a fundamental part of the human endeavour, and constitutes the essence of our perception of the world around us; it shapes the way we build knowledge and produce scientific thought.

Every mathematical entity *belongs* in a definite (mathematical) place; it “inhabits” a certain realm and not another. This classification system is important, since it avoids contradictory notions⁵. At any level of abstraction, these entities become unexpectedly mesmerising and intriguing; points, lines, paths, vectors, surfaces, volumes, and even space itself, are part of these entities. Three dimensional objects constitute, however, only a few particular cases of interest; one can eventually escape this “prison” of the third dimension and enter a wonderful realm in multiple dimensions, where the notions of *big* and *small*, or *far* and *close* are deprived of their usual meaning and the true nature of space is revealed, this is the world of calculus and topology.

Calculus is where we can closely examine the very large and the utterly small just for the sake of better understanding the space we live in (or the different spaces we might just happen to live in). Topology allows us to speculate about space itself, its shape, structure, and “texture”. This building method of mathematics is exactly what is meant with the notion of “going forward,” not in a sense of progress, but in the sense of a building process, i.e. we start up with a few concepts and explore where we, amongst a handful of logical rules, might just be *swept off to*. We explore a world that is different from our own physical world, which we explore by means of experimentation, and yet this abstract versions of space resemble our heuristic notions in at least some way, shape, or form.

By the end of the 19th century there was a widespread conviction that science, physics in particular, was an almost “complete” branch of human knowledge. Mathematics was thriving, and with the exception of a list of 23 problems⁶, it seemed as though it was about to be “completed” as a formal, absolute, consistent apparatus. David Hilbert (1862 - 1943), amongst other mathematicians like Henri Poincaré (1854 - 1912), felt otherwise.

As suggested by Hilbert in his speech addressing the Second International Congress of Mathematicians in Paris (1900), mathematics was supposed to face the beginning of the century with the bold optimism of solving a few problems that would eventually “complete the puzzle,” and converge into a perfect mathematical system.

The importance of mathematics in our conception of knowledge transcends any geographical and chronological boundaries; it transcends even the artificial boundaries between disciplines or different branches of knowledge, providing every piece of our formal understanding of the Universe with a means to convey truth and reason, even beyond the limits of humans’ lifespan. It is because of formalised knowledge that we have managed to possess historic and scientific consciousness. One of the 23 problems stated by Hilbert in his programme dealt precisely with the task of now providing mathematics with formal grounds; more precisely, it sought after a finite, concise proof that no contradiction could be

⁵ *Belonging* means, in a set-theoretical sense, that every mathematical object is an element amongst a collection of other elements of its kind, all contained in what is called a *set*

⁶ Hilbert’s speech at the 1900 International Congress of Mathematicians in Paris, France. The original speech (in German) can be found at: <https://www.math.uni-bielefeld.de/~kersten/hilbert/rede.html>

obtained in the formalism of mathematics⁷. Other than intuition, there was no formal way to know if a mathematical statement was correctly deduced, regardless if it was expressed in the proper mathematical language. Moreover, there was no well-defined mathematical language and no decisive way to distinguish a subtle error in a mathematical proof. In other words, there was no rigorous way to know if all the work in mathematics done so far was *consistent*.

Hilbert suggested that any formal theory that unambiguously provided mathematics with solid foundations would suffice as a method to prove its *consistency*. Since the development of mathematics during the 19th century had had a prolific growth, it was indispensable to guarantee that no two branches of mathematics would come up with (formally proved) results that could potentially contradict each other. If mathematics was indeed a foundational subject, what could it be formally based upon? Moreover, even if this question had an actual answer, the problem of providing this new subject with proper foundations would arise immediately, and the problem goes on indefinitely. Despite these difficulties, Hilbert thought that, as any other mathematical problem, the issue of consistency, i.e. of proving the consistency of the mathematical apparatus, actually *had* a solution, mathematicians had just not found it yet; it would, however, one day be found. As anyone would intuitively think, mathematical problems are supposed to be solved; for any question there is an answer, and these two always come in pairs. If a solution had not yet been found, new methods would be developed until an answer was reached.

Ernst Zermelo (1871 - 1953), and Abraham A. Fraenkl (1891 - 1965), amongst many other mathematicians, physicists, and philosophers, provided a possible solution to the problem. Since any idea consists of a series of underlying principles, they thought that a properly selected collection of basic, quintessential, and undeniable notions, the most primordial notions one could describe mathematics with, would suffice to hold the mathematical edifice together. They proposed a set of 14 independent axioms from which all of mathematics could be deduced. This idea prevailed, and it constitutes the essence of our current approach on mathematics, based on the conclusions one can logically derive from the set of axioms nowadays referred to as *ZFC*, i.e. Zermelo-Fraenkl-Choice⁸.

Axioms are starting points, statements whose validity one does not question. Changing such rules is valid, but it constitutes the invention of another (either different or equivalent) system. Notice that changing the axioms (rules) of an existent, consistent theory, does not imply that the new theory will still be consistent. Throughout history, we have had understanding of a vast diversity of concepts. With an axiomatic approach to mathematical concepts, it is clear that each concept's existence is either properly justified by these axioms, or deduced from them⁹. This means that any statement expressed in the language

⁷More on this topic can be found in: ZACH, R. "Hilbert's Program", *The Stanford Encyclopedia of Philosophy*, (Summer 2019 Edition), Edward N. Zalta (ed.)

⁸The *axiom of choice* deserves a separate text on its own. It formally states a way to choose elements from an infinite amount of sets, indistinguishable from one another.

⁹The rules of a game are axioms, since one does not question their validity. Take chess, for instance, and add a rule that states that pawns can no longer capture any piece. This will

of mathematics is either an axiom, or it can be logically inferred from one (or many) of these 14 statements. The set of axioms is of course independent, i.e. none of them can be deduced from one another¹⁰.

This approach provided some clarity in the search for foundations, but it did not solve the problem of consistency, and it came along a series of problematic consequences. Not everything deduced from these axioms stayed within the boundaries of human intuition. One of them, for example, states the primordial existence of an infinite set, i.e. a collection containing an infinite amount of elements, not potentially, but *actually*. This not only contradicts our intuitive (Aristotelian) notion of infinity being inherently *potential*, but from the existence of infinity as an actual (finished, i.e. not just potential) entity one can formally prove the existence of an infinite amount of *different infinities*, each one being strictly greater in size than the previous one¹¹.

This kind of counter-intuitive results obtained from set theory were not the only problem that arose from the Hilbert's Programme. The mere notions of *existence*, or *truth*, were problematic, since they had to be formally and unambiguously defined. A solution to many of these problems was being sought, and great efforts by the mathematical community were involved. However, one ontological misconception about the way we create knowledge is to think that problems constitute a biunivocal structure of *questions* and *answers*. To show that the actual apparatus we use to prove the logical validity of other sciences was itself logically valid, was some sort of self-referred paradox. This represented an inevitable necessity for a paradigm shift in rational thought, since some of the famous Hilbert's Problems can actually be proved to be *unsolvable*. It was, indeed, in 1931 that Kurt Gödel (1906 - 1978), an Austrian logician, formally proved that¹²

(·) If we choose a well defined set of axioms and inference rules in which only *true* statements can be deduced, there will *always* be statements, expressed in the formal language of the system, which are impossible to be proved, i.e. neither their veracity nor their falsehood can be proved.

(··) It is absolutely impossible to prove the consistency of a system within the system itself, i.e. one cannot deduce from the axioms that the axioms will not, eventually, exhibit any contradictions.¹³

lead to games that cannot go on after a few turns.

¹⁰After decades of development in *set theory*, some of the axioms were found to be a deduction from the others. Nowadays, there are only six axioms considered as a foundation of set theory. A lot more on this can be found in: JECH, T., (2003). *Set Theory*. Springer-Verlag Berlin Heidelberg. Springer Monographs in Mathematics.

¹¹CANTOR, G. (1874), "Ueber eine Eigenschaft des Inbegriffes aller reellen algebraischen Zahlen", *Journal für die Reine und Angewandte Mathematik*, 77

¹²GÖDEL, K. 1931, "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I", *Monatshefte für Mathematik und Physik*, v. 38 n. 1, pp. 173–198.

¹³For a detailed discussion on the subject, see the following article (in Spanish): TORRES, C. 2000, "La lógica matemática en el siglo XX", *Miscelánea Matemática SMM*, n. 31

The language of logic is a formalised version of the founding principles of reason. We trust knowledge because we think it to be *logic-proof*. The most basic, conceptual structures we need to understand any kind of logical reasoning, however, are based on the primal notions of sets. Take the concept of *two*, for instance; the only way to define formally what the number *two* is, is by thinking of all collections one can think of that contain exactly “two” elements. It is from the understanding of the *void*, i.e. the concept of nullity, of an empty set, that we understand unity, and it is with the combination of the existence of void and unity that we understand *two*. It seems then, that logic, which is the study we undergo to comprehend our own mental structure, our language, our perception of reality, and most importantly *truth*, has a very intricate relation to set theory. Both of them seem to inspect and scrutinise the other, support their validity; they seem mutually to provide a justification for the other’s existence, as arrows that aim at each other.

Finally, it is through drawings, diagrams, or equations that we can grasp the ethereal fragility of this infinite art. Just as poetry captures the essence of language, equations withhold the essence of reason. In a non-trivial fashion we could analyse the role played by *depiction* in the poetry of mathematical reasoning. This leads to an inevitable third path taken perhaps by wonder through the intricately complex realm of reason. There exists no way to convey such elegance if it is not through pictorial representation. Geometry plays a quintessential role in this process, one which is always intertwined with creativity in order to merge it together with reason, however different the manifestations may be.

Amongst many other comments one could assert, it is appropriate to identify geometry as a formal representation of the levels of abstraction reached with calculus, algebra, or any other branch of mathematics. More importantly, the visualisation part of the learning process defines and shapes the knowledge we are building in many different levels. The physical sciences, for example, are notably affected by the *kind* of geometry they are based upon.

Our first encounter with the physical world comes along a Socratic perception of the world; we perceive a three-dimensional flat space, a 3-D grid of 90° cubes that seem to contain the wholeness of physical reality. Even when we expand this spatial perception into that of a curved, spherical Earth, we still imagine it to be embedded in a sort of “rectangular” Universe. The more we develop our physical understanding of this Universe, we change from this Socratic vision, where things remain in their standing states, and try to come back to it, unless affected by an external force, to a Newtonian perception of reality, where movement is relative, and things actually remain on a constant, rectilinear-motion-state until an external force acts upon them. In any case, Euclidean geometry *shapes* the way we perceive reality.

By the end of the 19th century, it started to seem clear how such a geometrical perception lacked the adequate elements to describe reality properly. Einstein’s proposal of a general-relativistic Universe involves much more than a conceptual paradigm-shift in physics; its true sustenance relies on the change of geometrical foundations. One cannot understand relativistic mechanics unless one accepts

the idea of a curved, hyperbolic Universe, where light travels not in straight lines as understood classically, but through hyperbolae, connecting space in the most efficient way, and allowing trully fast objects to travel in time into the future.

It is in this exact same fashion that geometry shapes quantum mechanics, providing it with its characteristic notions. Some of the underlying physical laws that were presented as fundamental for quantum theory are well summarised in Heisenberg's relations; their geometrical implications are no less than astounding, but fairly sophisticated for an introductory version. These inequalities define a non-commutative geometry, an infinite-dimensional space where symmetry is encoded by operations whose order *does* actually influence the results of our measurements, and where these operations are closely related to the notions of harmony in music.

When properly developed, geometry can be a formalised path to depict the *truth* of scientific endeavour. Either by pictorial representation, formalisation, or the philosophy of the mathematical language, the means to reflect the massive edifice of knowledge and the structure of reason continue to develop, and more paths are found that lead us into a spiral of infinite conundrums.

Chapter 1

Introductory Topics

§1 Our Perception of Matter and the History of the Atomic Model

†† Throughout the centuries, our perception of the physical world around us has evolved; we have developed ever more complex experimental techniques to explore the way our Universe works. The first inklings of an atomic theory of matter came with the ideas of Leucippus (c. 5th cent. BCE) and his pupil, Democritus (c. 460 BCE - c. 370 BCE). They proposed that matter is composed of indivisible particles. Although this was a revolutionary perspective at the time, merely by denying divine intervention in design¹, or simply by suggesting a “primitive” quantisation of matter or space, one can easily identify how this atomic theory differs from our current understanding of matter.

§1.1 Atomism

Atoms, to begin with, were supposed to be intrinsically unchangeable, had different shapes and sizes, where different kinds of atoms resulted in different textures, tastes and colours. Men, for example, were supposed to be made of *men-atoms*, women of *women-atoms*, wood was supposed to be composed of *wood-atoms*, and so forth. Atoms were essentially a solution to a metaphysical problem, finding the *origin*, or *first principle* of all things that exist (the $\alpha\rho\chi\eta'$). We must be cautious if we decide to include Ancient Atomism when studying the modern atomic theories. It should suffice to say that, despite the importance of the first atomist conceptions of the Universe, one should simply be careful in any attempt to organise knowledge, for a simple inaccuracy may lead to grossly misleading interpretations. The pre-Socratic perspective mostly comes from the idea that cutting something in half enough times leads eventually to an “atomic” level, i.e. a moment when the physical matter being cut becomes indivisible. At this point one finds the constituent elements of matter.

¹BERRYMAN, S. “Ancient Atomism” *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.)

†† However intelligent these assumptions were, especially at a time when no previous scientific work or accumulated knowledge was available, they were merely *blind* guesses of what the nature of physical reality *might* be. Gradually, rigorous observation and experimental data was accumulated, allowing us to come up with formal models of what matter might look like at its innermost levels. These models have been refined over time as our technological capabilities evolve and we become able to study the structure of matter at increasingly deeper levels, each generation of scientists building up over the foundations left by the previous generation.

The atomist ideas were then left aside, and only a handful of scientists kept them alive during the Middle Ages and the Renaissance. Chemists of the 18th century, almost 2,200 years after the first atomist ideas, studied the composition of different materials, and noticed that most of them were merely a combination of other, more simple and pure, substances. These pure substances were composed of *elements* or combinations of such elements called *compounds*. Compounds, e.g. water, could be broken into their constituent elements, and they would always have the same proportions.

§1.2 Dalton

By the end of the century, chemists had already established these observations as a physical law. A fixed amount m of a compound C that could be decomposed into m_A grams of A and m_B grams of B would always be such that the proportions of these masses are conserved. I.e. the corresponding masses of the products will *always* have the same proportions, regardless of the original amount. This fixed ratio is referred to as the *law of definite proportions*. For example, 8 grams of water are always composed of 11.11% (1 g) of hydrogen and 88.88% (8 g) of oxygen.

Mass, as one could expect, is conserved in a chemical process, i.e. a process where matter changes its structure, its shape and/or form. It means that matter cannot be created or destroyed, i.e. *the total mass in a chemical reaction must remain constant*. This conservation law was proposed by Antoine Lavoisier (1743 - 1794), a French chemist, but is also attributed to Mikhail Lomonosov (1711 - 1765), a Russian scientist and writer from the 18th century.

A refinement to these two laws was proposed by the British chemist John Dalton (1766 - 1844) in 1808. His statement reads:

If two elements form more than one compound between them, then the ratios of the masses of the second element which combine with a fixed mass of the first element will be ratios of small whole numbers. As an example, consider

a fixed amount of carbon reacting with oxygen to form an oxide, a reaction well known by chemists of the time. The product would always be one of the two following: CO or CO₂ (carbon monoxide, and carbon dioxide, in modern notation). After a decomposition reaction, the corresponding masses of oxygen in the two compounds that combine with this fixed mass of carbon should be,

according to the law of multiple proportions, in a whole-number ratio. So, in 100 g of the first compound there are 57.1g of oxygen and 42.9g of carbon. This means that the mass of oxygen relative to the mass of carbon is:

$$\frac{57.1}{42.9} = 1.33 \quad \frac{\text{g of oxygen}}{\text{g of carbon}}$$

I.e. 1.33 grams of oxygen per gram of carbon. Accordingly, in 100 g of the second compound, there are 72.7g of oxygen and 27.3g of carbon. The relation between the mass of oxygen and that of carbon is:

$$\frac{72.7}{27.3} = 2.66 \quad \frac{\text{g of oxygen}}{\text{g of carbon}}$$

I.e. 2.66 grams of oxygen per gram of carbon. Comparing the mass of oxygen per gram of carbon of the second oxide with that of the first one, we obtain the following ratio:

$$2.66/1.33 = 2$$

I.e. the masses of oxygen that combine with carbon are always in a 2:1 ratio, thus following the law of multiple proportions. This empirical evidence showed chemists of the time how matter *behaves*; Dalton's interpretation of these observations was that these oxides consist of one (and two, respectively) oxygen *atom(s)* joined to a carbon atom. Dalton acknowledged that an atomic theory of matter would explain these and other physical observations properly. He thus theorised that such model should obey the following axioms:

- (·) Elements are composed of minuscule particles called *atoms*.
- (··) All atoms of a given element are identical with respect to physical properties, and thus indistinguishable from one another.
- (···) Atoms of a given element are different with respect to all physical properties from atoms of another element.
- (····) Atoms involved in a chemical reaction unite chemically in simple numerical proportions (e.g., 1:2, 1:3, 2:3 etc.) to form *compound atoms* which are now known as molecules. A given compound will *always* have the same relative number and type of atoms.
- (–) Atoms are indivisible, and cannot be sub-divided by means of any physical or chemical process. A chemical reaction simply changes the way these atoms are grouped together.

Dalton's atomic model was useful to explain all the previous observations and to predict the way (or different ways) atoms could combine themselves to create chemical compounds. Although this atomic model was not widely accepted at first, it was eventually recognised and approved as a correct way to explain physical and chemical processes.

§1.3 J.J. Thomson

By the beginning of the 19th century, scientists had a much better understanding of the behaviour of atoms. Such an understanding could explain and even predict how atoms bond to create molecules, but it did not answer the fundamental question *why?* Why do atoms behave the way they do? To answer this question one should ponder upon the internal structure (if any) of atoms, something that was prohibited by Dalton's axioms.

A series of experiments with cathode rays had shed some light on what might just be the answer to this question. A beam of cathode rays is produced inside a tube with the aid of an electric current, flowing from a piece of a negatively charged metal (cathode) to a piece of a positively charged metal (anode)². Back in 1654, the German scientist Otto von Guericke (1602 - 1686) invented the vacuum pump, allowing physicists to experiment with high voltage electricity travelling through *low-density* (rarefied) air. By the mid 1800's it was well known that electricity could produce glow inside a glass tube partially evacuated of air.

It was not clear if these cathode rays were immaterial, like light, or "*in fact wholly material, and [...] mark the paths of particles of matter charged with negative electricity,*"³ to quote the British physicist Joseph John Thomson (1856 - 1940). Thomson first noticed that cathode rays travelled in straight lines; he then noticed that their path could be bent (deflected) by the presence of an electric field. By placing two charged plates around the glass tube, he noticed that the beam was always bent towards the positively charged plate. This suggested that the beam was negatively charged.

J. J. Thomson then confirmed this hypothesis by bending the beam of cathode rays with a magnet. By measuring the heat generated when these rays hit a thermal attachment, and comparing it to their magnetic deflection, he managed to estimate the *mass* of the beams, if in fact these beams were actually material. In that case, cathode rays must be composed of something that is negatively charged. Its constituent elements must be about one thousand times smaller and lighter than hydrogen atoms, which was already the lightest element known. Finally, Thomson noticed that all metals could be used as anodes and cathodes to produce cathode rays and these were always the same, they deflected the exact same amount, and the hypothetical mass was always the same.

All these observations contributed to J. J. Thomson being the first one to suggest, in 1897, the existence of a sub-atomic, negatively charged, particle. According to his idea, all elements were composed of a positively charged mass that counteracts the negatively charged particles scattered randomly throughout the inside of the atom. He called these sub-atomic particles *corpuscles* and suggested a new atomic model which he called the *plum-pudding* model of the atom, reminiscent of the British dessert, with raisins randomly dispersed within

²Positively charged cations always move towards the cathode (hence their name) and negatively charged anions move away from it. Reference: "Cathode" at *Wikipedia, The Free Encyclopedia* 2016

³THOMSON, J. J. (1897). *Cathode Rays*. *Philosophical Magazine*. 5. 44 (269): 293.

the pudding. The name would eventually change into the more familiar *electron*.

According to this perspective, the low pressure air inside the tube consists of atoms of a particular gas, and there is enough space between the gas atoms so that electrons can accelerate to produce the cathode rays inside the glass tube.

§1.4 Rutherford

The use of cathode rays and the possibility to experiment with photoluminescent materials⁴ gave physicists of the 19th century a great insight on the study of the internal structure of matter. At the time, certain (radioactive) elements were known to emit what Oceanian physicist Ernest Rutherford (1871 - 1937), a former student of J. J. Thomson, called *alpha rays*. Elements like uranium and radium spontaneously and unpredictably emit this radiation in the form of a particle, also called an *alpha particle*; nowadays such particles are properly identified as helium nuclei.

Back in 1908, Rutherford was studying this phenomenon by trying to measure their charge and mass. Since alpha particles are obviously too small to be detected by a microscope, the way to study them was by measuring their deflection in a gas chamber implemented with an electric field. Gas particles were ionised by the radiation (alpha-particles) upon collision, which would in turn deviate their path slightly. A photosensitive material could be set, surrounding the chamber, to detect the deflected alpha-particles.

There is no such thing as solid matter in the atomic realm; this means that any detectable deflection should be caused by an electric field. At a macroscopic scale, matter bounces off surfaces, like a marble against the wall, but alpha-particles are able to pass through a thick piece of matter without a significant deviation. One would expect from Thomson's plum-pudding model that atoms have a weak electric field, unable to deflect significantly (no more than a few degrees) a beam of alpha-particles. However, Rutherford and two of his students (and later colleagues), Hans Geiger (1882 – 1945) and Ernest Marsden (1889 – 1970), found a portion of alpha-particles being shot at a thin gold foil was being deflected by angles close to 90°, *almost as incredible as if you fired a 15-inch shell⁵ at a piece of tissue paper and it came back and hit you*, in Rutherford's own words.

The positive charge within the Thomson atom is homogeneously spread out over the atom's entire volume, and electrons are randomly scattered throughout the positively charged background. The large deviation of alpha-particles led Rutherford to think that the positive charge in the atom is not spread out over the atom, but in fact heavily concentrated at its centre. According to Coulomb's Law, which describes the interaction between charged particles, electric force is inversely proportional to the square of the distance between such particles; i.e. the closer a charged object is to another charge, the stronger the force, more so given the fact that the distance should be squared.

⁴For instance phosphorescent and fluorescent matter, i.e. matter that can absorb electromagnetic radiation in the form of light and then re-emit it over a period of time

⁵a cannonball

Coulomb's law:
$$F = \frac{K \cdot Q_1 Q_2}{r^2}$$

where F is the force, K a constant, Q_1 and Q_2 the corresponding charges, and r is the distance between the charged particles or objects. Mathematically, if these two particles were to be arbitrarily close together (the distance r being arbitrarily close to zero), the force would be infinitely large.

He called this charged cluster a *nucleus*, and he concluded that electrons should be orbiting around this nucleus, introducing the first non-static model of an atom, precisely the *planetary* model many of us are used to. In that case, positively charged particles, like alpha-particles, close to the nucleus would feel a strong electric force repelling them. This explains the large deviation angles constantly measured when scattering alpha-particles.

Later, Rutherford found that a nitrogen atom could eject a hydrogen nucleus when hit with alpha particles. This led him to conclude that these hydrogen nuclei *must* be a constituent part of nitrogen atoms, and in fact of all atoms in general, a building block of matter nowadays called a *proton*. It is worth to notice that the mere idea that radioactivity (alpha-radiation) was in fact an atomic phenomenon, i.e. something that involved a change in the internal structure of the atom, was revolutionary per se.

§1.5 Bohr

Rutherford's model was very successful; it allowed him to discover the proton and to understand the internal structure of the atomic nucleus itself. Further research made it possible to count the electrons around an atom, and the neutron was also discovered in 1931-32 (first thought to be a proton-electron neutral combination in the nucleus). There was, however, a substantial theoretical problem with Rutherford's atomic model.

An accelerated charge emits electromagnetic radiation, it therefore radiates a certain amount of energy that is proportional to the square of its acceleration. This means that an electron revolving around the nucleus of an atom would inevitably lose energy due to centripetal acceleration until it collided into the nucleus. This would happen in a fraction of a second, rendering the existence of atoms completely impossible. This meant that Rutherford's planetary model must be incorrect.

Niels Bohr (1885 – 1962), a Danish physicist, simply decided to *postulate* a solution to the problem, as opposed to the usual experimental approach. He took previous observations into consideration and refined Rutherford's model with the following principles:

(·) An atom possesses stationary orbits; to be precise, electrons revolving around the nucleus do so in certain specific orbits in which they do not radiate energy. These orbits correspond to fixed values of energy. Changes in the atom's energy are due to electron transitions between such orbits.

(··) Absorbed or emitted radiation during any transition between stationary

states of energy E_1 and E_2 is monochromatic, i.e. it consists of a single frequency given by the following relation:

$$\Delta E = h\nu$$

where $\Delta E = E_2 - E_1$, h is a physical constant, ν is the radiation's frequency, and $E_2 > E_1$.

One can see how axiom (\cdot) is in perfect agreement with Planck's and Einstein's famous relation; it also implies that electrons **cannot** have any intermediate orbits, i.e. between two consecutive, discrete orbits⁶. Thus an electron in transition from one orbit to the other appears to change instantaneously, hence the expression *quantum leap* or *quantum jump*. Energy is either absorbed or emitted in quantised amounts, corresponding to such "jumps;" furthermore, the emitted radiation has a frequency ν that is equal to the electron's orbital frequency. Finally, there is a lowest possible energy state, and it provides the smallest possible orbit. From this, one can compute the radius of such orbit to be $r_{Bohr} = 5.29 \times 10^{-11}m$. Bohr's atomic model was the first to include a quantum description.

One of the great successes of Bohr's model was its capability to explain the emission spectrum of hydrogen, something no previous theory could do. Back when Isaac Newton coined the term *spectrum* to describe the decomposition of white light into its constituent colours, scientists began to improve the optical mechanisms to study light. Subsequent experiments involved not just sunlight, but different light sources as flames, and even other stars. By burning different elements, scientists discovered that each element has its own, characteristic radiation spectrum.

By the end of the 19th century, physicists were using glass tubes to study incandescent gases of various elements at different temperatures; the light they emitted was passed through a prism, and decomposed to study its constituent wavelengths. Instead of a rainbow, each element seemed to be composed of its own series of spectral lines, often called *Fraunhofer lines*⁷. These lines are each element's *fingerprint*, allowing scientists to identify, for example, the constituent elements of stars and planets just by analysing the light we receive from them.

Spectral lines are perfectly explained by Bohr's model as an electron's transition from one energy state to another, the line's frequency given by the specific energy levels the transition takes place in. Moreover, using Bohr's model, physicists were able to predict series of lines in the hydrogen's spectrum that are out of the range of visible light.

§1.6 Schrödinger

Elements as helium, lithium, etc, also have their own characteristic spectral lines. Bohr's atomic model, however, cannot explain the spectral series of any

⁶These concepts are properly defined and discussed in sections § 2.5 and § 3.1

⁷honouring the German physicist Joseph von Fraunhofer (1787–1826), who first identified these as absorption spectral lines in sunlight.

element beyond hydrogen. A few attempts were made to refine Bohr's model, but they were all unsuccessful or extremely complicated. It was revolutionary, since it gave birth to a *primitive* quantum theory, but it still had classical elements that made it insufficient. A new perspective seemed necessary in the study of atomic mechanics, and this new theory should incorporate the quantum nature of atoms as its founding principle.

The Bohr model was eventually replaced by quantum mechanics. Some fundamental changes were made; the electron, for instance, is described as a *wave-function* instead of as a particle; it occupies an atomic orbital with a given probability rather than following a trajectory in an orbit. Even though the allowed energy levels of the hydrogen atom remained the same as in Bohr's model, quantum theory goes way beyond the classical view and constitutes a major paradigm shift in physics, a *quantum leap* if you will.

Schrödinger's equation is, undoubtedly, a centrepiece of quantum theory. It was published in 1926, and it describes *quantum states* rather than localised particles and well-defined paths. With quantum mechanics, many of the spectral phenomena that Bohr's model failed to explain were finally and correctly clarified.

Most of its elegance and beauty are to be found in its mathematical descriptions; it is a vast subject, and it has enlightened our exploration of the Universe at its most fundamental scale. This is, however, an ongoing quest; it is by no means finished, and as our understanding of the *inconceivable nature of Nature* progresses, the range of questions we can ask, as well as the amount of knowledge waiting to be discovered, widens just as rapidly.

§2 The Double Slit Experiment: A Paradigmatical Anomaly

†† We begin by discussing the two familiar notions of *matter* and *waves*. Intuitively, matter is the quintessential constituent of everything; it provides objects with *bulkiness* and mass. We relate waves to oscillatory motion; *something* moves and creates waves, these transfer energy (as in the waves of the ocean) and bring information along with their oscillations (as in radio or television waves). The intricate differences and the surprising relations between bulk matter and periodic oscillation lie precisely at the core of the *quantum* nature of our Universe.

†† This distinction between matter and waves should be made carefully and in close detail; matter can be localised in a bounded region of space, waves are spread out and occur across the wholeness of space. However, it might not be obvious how or why one should make this seemingly narrow-minded observation in the first place; after all, matter and waves are not even close enough in the macroscopic world as to produce any kind of confusion. When we think about it further, it is not even clear why someone *had* to clarify this to begin with.

The story begins in the late 17th century, when a most avid discussion was held amongst the great physicists and philosophers of the time. Some of them, as Isaac Newton and René Descartes, believed light consisted of minute corpuscles

which travelled through space in straight lines and interacted with matter either by being absorbed or by being reflected. This view can explain why light bounces off surfaces according to reflection laws; it cannot, however, explain some other physical observations as refraction or polarisation.

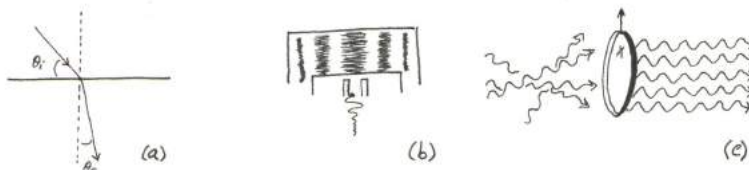


Figure 1.1: Refraction (a), diffraction (b), and polarisation (c) of light

When light is travelling from one medium to another, like air and water, its path gets angled. Its velocity in the first medium is not the same as it is in the second one; by changing its travelling direction, light travels through the path that optimises time. This can be easily seen with a spoon in a glass of water; from afar, it seems as if the spoon were broken into two separate pieces at the surface of the water. This phenomenon is called *refraction*, as shown in Figure 1.1 (a).

When passing through a crystalline structure, like the wings of a butterfly, light goes through the tiny slits between the scales of its wings. The reflected light within the wing's structure is usually out of phase, meaning it is misaligned and thus cancels itself out partially. Blue light, however, comes out perfectly aligned, thus resulting in the beautiful blue patterns we see. The notion of "alignment" can be better explained by understanding light as an oscillatory phenomenon. In contrast to any other colour in nature, blue does not come from pigments, but rather from this kind of optical phenomena; this is called *diffraction*, as shown in Figure 1.1 (b).

Light can also pass through a grid of aligned molecules that only allow "parallel" beams to pass through, like camera lenses' filters; the result is a stream of *polarised* light, as is shown in Figure 1.1 (c), another phenomenon that can only be explained by means of periodic oscillations, i.e light waves.

The corpuscular view of light, however, prevailed for more than a century, and it was perhaps due to certain sociopolitical accounts, at least partly. Distinctively, Newton's prestige was unprecedented; he was undoubtedly one of the most acclaimed scientists of the century, and his views on any subject were inevitably something people thought worth taking into consideration. Christian Huygens, a dutch contemporary of Isaac Newton for example, and Robert Hooke, also an English natural philosopher, proposed a wave theory of light, but this perspective was partly undermined by the scientific community.

§2.1 The Problem with Light

The *wave theory of light* compared propagation of light to the motion of waves in water, but with some sort of *luminiferous ether*⁸ as a medium of propagation. This predicted the phenomenon of interference, something that was already known to scientists from experiments with sound; it also suggested a correlation between colours and wavelengths. One later experiment by Thomas Young (1773 - 1829), an English polymath, proved how light did manifest a wave-like behaviour under the proper circumstances, with an experiment that would turn out to be decisive, even for the understanding of quantum physics.

†† For almost a century, the empirical evidence of Young’s experiment prevailed as an experimental account for the wave theory of light. The problem came when, at the beginning of the 20th century, another decisive experiment proved an undeniable matter-like behaviour of light. This “undeniability” of both facts, light being a particle, and light being a wave, has ever since remained an enigma for modern physics. The most simple and elegant way out of the problem is simply to accept both facts as true, and consider light to be a stream of particles whenever we work with experiments of the second type, and a set of electromagnetic radiation every time we perform experiments of the first type: whatever suits our scientific goals.

§2.2 What we Understand as Matter

In accordance with Democritus’ ideas⁹, we understand matter to be composed of tiny *ατομοι* (from *ατομος*), indivisible particles; these, as all matter, can be located as “lumps” in space. Atoms constitute the inherent structure of matter. As obvious as this idea may sound, it is not trivial, and much less was it at the time.

One must, however, be very careful when stating how or why Democritus postulated this idea. It would be an anachronistic inaccuracy to suggest that his ideas were in any sense equivalent to our present understanding of the atomic theory. Moreover, it would be rather irresponsible to suggest that he, or Epicurus whilst revising the atomist perspective, had in some way foreseen the existence of atoms as a scientific fact. As was said before, it was only a philosophical concept, an idea to explain the origins of matter.

The atomist ideas were mostly forgotten in Europe until the late Middle Ages, when some texts by Aristotle were rediscovered. Empirical evidence that support the actual existence of atoms was not found until the second half of the 19th century. Way before that, at the beginning of the century, chemists had already discovered a wide variety of chemical elements of which matter is

⁸Waves are propagated in a medium, water for instance, but light did not seem to need a medium to do so. It was then thought that space itself was submerged in a yet undetected medium where light was able to travel. Later experiments, namely that of Michelson and Morley (1887) proved this wrong.

⁹One of the best references in philological terms, inter alia, is the *Stanford Encyclopaedia of Philosophy* (SEP). More on Democritus and atomism can be found at: BERRYMAN, S. “Democritus” *The Stanford Encyclopedia of Philosophy* (2016 Edition), Edward N. Zalta (ed.)

composed. As was briefly discussed in §1, Joseph Proust's (1754 - 1826) observations on chemical compounds breaking down into their constituent elements led to the *law of definite proportions*, also influenced by Antoine Lavoisier's work and corroboration of the *law of conservation of mass*. Later work by the English chemist John Dalton showed that such examples of fixed patterns, i.e. definite and multiple combining proportions in chemical reactions, could be well and elegantly explained by an atomic theory of matter.

It was not without great effort and trouble that the scientific community of the time accepted the atomic theory. As an example, take Ludwig Boltzmann's (1844 - 1906) contributions to the understanding of fundamental, physical concepts such as temperature. Along with other noteworthy thermodynamicists of the time, Boltzmann established the bases of classical statistical mechanics. Despite such noteworthy contributions, many scientific journals refused to allow him to refer to atoms as real entities; instead, they were to be understood as mere theoretical concoctions. It was only after experimental confirmations at the beginning of the 20th century that the atomic theory of matter gained enough credibility and empirical evidence as to solidify its theoretical foundations.

From a certain perspective, everything in the macroscopic world can be seen as a simple, undivided *lump* of matter. For instance, seen from afar (or far enough at least), a baseball, a mug, or even a chair can resemble an elementary particle, so we set on a quest to decompose matter on ever smaller and smaller constituent particles. The structure of such particles remains unknown until better experimental methods and equipment are developed, and in the meantime it suffices to imagine any atomic or subatomic particle as a tiny sphere (a sphere being the most symmetric geometrical solid, thus a more convenient form than a cube, for example). That being the case, we focus our attention to the behaviour of *tiny spheres*, and assume that the microscopic world can be explained in terms of known particles, like protons, neutrons, electrons, quarks, and so forth.

There is an almost obvious question that arises when studying the world around us. In order to undergo a study of the structure of matter, one needs to explore its internal components; to understand how they work, one often requires an internal view of the objects of interest. In much simpler words, we need to *break* stuff to see how it works. As rudimentary as this idea may sound, it is the basis of many different projects in today's research. CERN colliders, for instance, fragment accelerated particles ("atomise" them, if you will) and classify the remains according to energy differences and their associated masses (among other physical properties), thus identifying them as new or already known particles.

In order to study the behaviour of matter in different situations, let us imagine a cannon-like device that can shoot lumps of matter, these being atomic particles (like electrons, for instance), or whatever "chunk" of matter one wishes to analyse. Everything is shot onto a screen, where it is collected, and the exact location of impact is registered. A photograph is a nice example of this kind of devices; light (whether as particles or as electromagnetic radiation, depending on our perspective) comes through a lens or pinhole, into the camera, and lands

on a piece of photosensitive material (a plate or film); it is later activated by silver nitrate, and registered as an image.

If we placed, between the shooting device and the screen, a barrier with two slits on it, then matter would bounce off of it, and just a fraction would be able to cross through the slits, all the way onto the screen. After a while we could remove the screen, and we would see a pattern of two straight lines. This should come as no surprise, since particles can either bounce off the barrier, or go through the slits; it is exactly what we would expect. Since everything we can equate to *lumps* in space (like particles) will behave accordingly, let us, for now, define *matter* as everything that, when collectively and progressively shot through the two-slit barrier, replicates the two stripes pattern¹⁰.

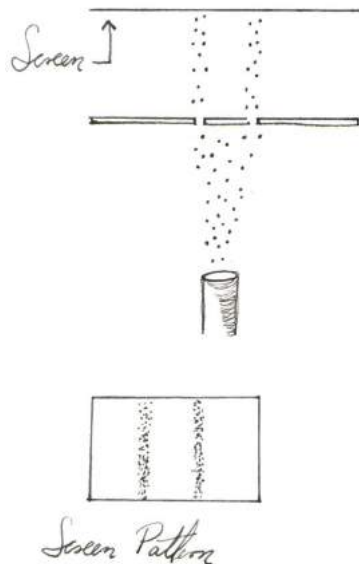


Figure 1.2: Particle scattering

§2.3 What we Understand as Waves

On the other hand, we could perform a very similar experiment in a container of water. This time, instead of *shooting* particles at the two-slits barrier, we locate a small object at a certain distance from the barrier, and induce the propagation of waves in the container by moving the object up and down. This oscillatory motion would produce a chain of water-waves as concentric circles around the

¹⁰The double-slit experiment is one of the archetypical experiments that exhibit a quantum behaviour. See the discussion found on TIPLER, P. A. (1970) *Foundations of Modern Physics*. Worth Publishers, Inc. Second edition. and BEISER, A. (2003) *Concepts of Modern Physics*. McGraw-Hill. Sixth edition.

object, which would propagate throughout the water, across the barrier, and onto the screen.

Whilst matter (lumps) either bounced off, or went through either one of the two slits, waves would create new concentric circles at each of the slits. Notice how the new propagation points replicate the oscillatory motion of an object located exactly at each of the slits. Once through the slits, the disturbances along the surface of the water would travel to the screen. Before we undergo a study on the components of waves in water or any particular medium, compare the approach and the technical procedures of this and the former experiment (depicted in Figures 1.2 and 1.4).

As a result, some of the disturbances in the water (coming from both slits, say, one from the left slit and one from the right slit) would add up to produce a larger disturbance, and some of them would cancel each other out, depending on the location on the screen. This is a phenomenon called *wave interference*. The outcome is a pattern on the screen, much different from the two stripes of the previous experiment, known as an interference pattern, manifested as a succession of stripes, much denser at the center and dimmer at the sides (“Screen Pattern” in Figure 1.4).

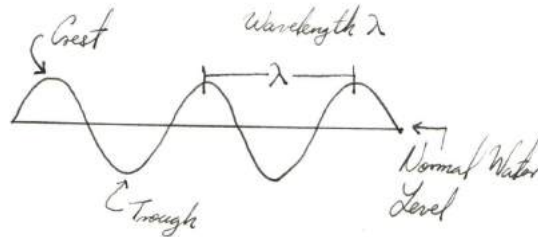


Figure 1.3: Waves consist of travelling “disturbances.” They have a wavelength λ , that corresponds to the distance between crests (or troughs), an amplitude A , that corresponds to the wave’s “height,” and a frequency, labelled ν , which corresponds to the amount of events (disturbances) that occur at a given point in space per unit of time.

The phenomenon of wave interference across a slit barrier which creates this palette of stripes is called *diffraction*. In the example of a butterfly’s wings, light is diffracted within the orderly structured pattern of the wing, at a microscopic scale. Only waves manifest diffraction, so let us say that everything that, when “shot” through the two-slit barrier, reveals an interference pattern is a *wave*. In a few words, waves are not *things* themselves, they are disturbances that, without manifestly moving (carrying) matter from one place to the other, can transfer energy. For instance, if two separate people hold a string (no matter how far they are) they can transfer a “message” of energy disturbance without moving, by using a *wave*.

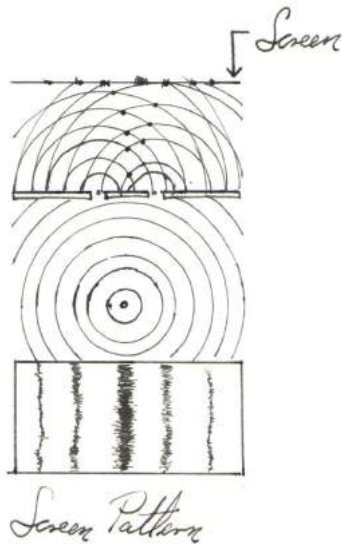


Figure 1.4: Waves in the two-slit experiment

§2.4 Light as a Wave

The two-slit experiment was conceived in 1800 by Thomas Young to prove that light has a wave-like nature. Light is then composed of waves with different wavelengths, each range of such wavelengths associated to a particular colour, violet light being that of shortest wavelength, followed by blue light, green, yellow, orange, and finally red light, with the largest wavelength. The refraction of white light by means of a crystal prism, which has become a canonical demonstration in many laboratories at high-school or undergraduate level, exhibits this orderly decomposition of light in a spectrum of colours and hues. Since different wavelengths get refracted by a different angle, i.e. the smaller the wavelength, the greater the angle of deviation, the components of white light are thus exhibited with this simple experiment. (Figure 1.5)

During the course of the same year, William Herschel (1738 - 1822), a German-British astronomer and musician¹¹, discovered the existence of light beyond our visual boundaries. He had previously discovered an increase in temperature whilst studying light with a red filter; he wondered if different colours could carry different amounts of thermal energy with them, and designed the following experimental setup: he passed light through a prism to decompose it as a rainbow, into its constituent ingredients; he then placed a thermometer at each of the different colours. He noticed that violet and blue light produced the least increase in temperature, whilst red light produced the greatest. His control

¹¹Not to be confused with his son, John William Herschel. They were both outstanding men of science and arts, the former's compositions having been praised even by Mozart and Beethoven, and the latter's work in astronomy being also of great importance.

thermometer, however, was resting next to the others, outside of the range of light's decomposition; to be precise, it was right next to the one measuring red light. He then observed an even greater increase in temperature on this far end of the spectrum, thus concluding the existence of light after the red limit of the visible spectrum; he had discovered infrared radiation, as it often happens in science, quite by accident.

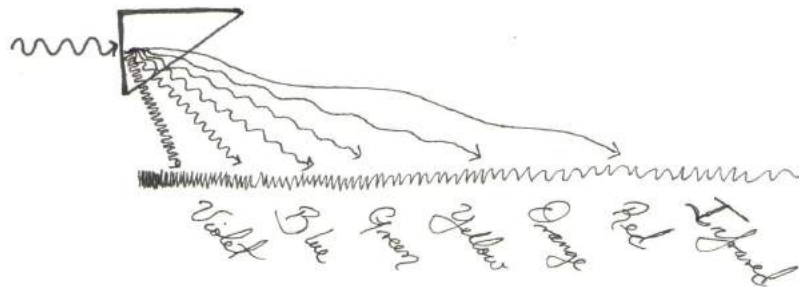


Figure 1.5: The prism experiment. The angle of deviation depends on the light's wavelength.

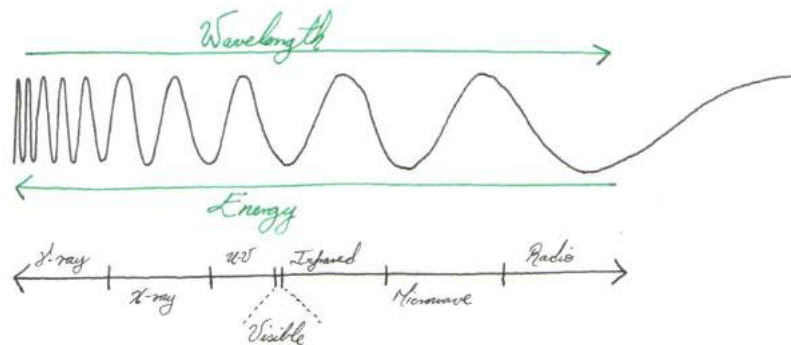


Figure 1.6: The electromagnetic spectrum. We now know that light consists of this whole range of wavelengths, only 4% of which humans are able to see. Some animals, like bats or snakes, are sensitive to wavelengths in other ranges of the spectrum. In terms of energy and frequency, any radiation immediately before the red side of the spectrum is called *infrared* light; any radiation immediately after the violet side is called *ultraviolet* light.

§2.5 Light as a Particle

By the end of the 19th century, back when most physicists and mathematicians believed science was about to be completed, there were only a handful of problems that seemed to be causing a delay, namely those involving certain phenomena with light. Other than that, physics seemed a complete description of Nature. After all, every one of its branches had already been explained in terms of a fully satisfying mathematical theory (Maxwell's equations had summarised electromagnetism, thermodynamics had had an astonishing success after the industrial revolution, and Newton's mechanics had fully described the known world, connecting the simplest motion at an almost microscopic scale to the complicated mechanics of planets, galaxies, and all other celestial bodies). One of these still unexplained phenomena was that of electric currents induced by light of specific colours shed on the surface of different metal plates.

Some of the experiments performed to understand this phenomenon eventually led to the creation of a very “primitive” quantum description of light. As was mentioned in §1, cathode ray tubes were used to study the internal structure of matter. J. J. Thomson's description of these rays as a stream of negatively charged *corpuscles* had proved to be essential to the following atomic models; a voltage (e.g. a battery) was used to create a current inside the glass tubes, but an external source of light could be used instead to trigger this electron flow by shining it onto one of the metal plates inside the tube. This is called the *photoelectric effect*¹². See Figures 1.7-1.9.

Back then, physicists had already identified this very particular situation in which electromagnetic radiation was not *behaving* as they expected. When radiated with light of different wavelengths, i.e. different colours¹³, a conducting material (like copper) manifested a flow of electric current from its surface, one which depended only on the colour. Blue and ultraviolet light caused a remarkably measurable current; red light, however, caused a much weaker electric current, and regular white light seemed to cause no current at all. The electrons on the surface “acquired” energy from the beam of light and transformed it into kinetic energy, thus creating the electric current. The dimmest radiation of ultraviolet light caused an electric flow.

¹²The photoelectric effect is one of the archetypical experiments that exhibit a quantum behaviour. See discussion found on TIPLER, P. A. (1970) *Foundations of Modern Physics*. Worth Publishers, Inc. Second edition. and BEISER, A. (2003) *Concepts of Modern Physics*. McGraw-Hill. Sixth edition.

¹³As one should recall from the explanation of the prism experiments in the 18th century, light is composed of a *rainbow* (quite literally) of different hues, each one corresponding to a colour. For a deeper explanation of the subject, a review of both the experiments of Newton and Herschel can be made.

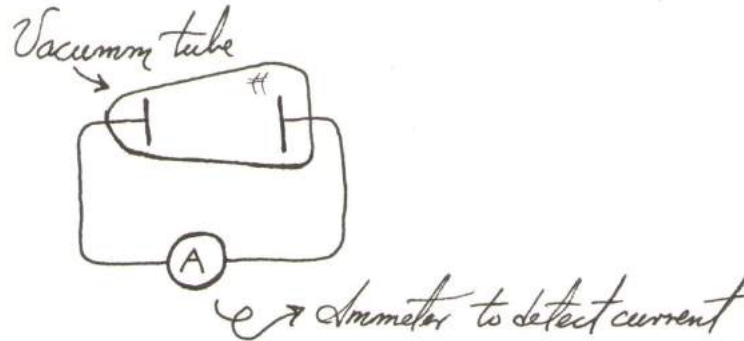


Figure 1.7: Experimental setup. Light is shed onto the metal plate on the right, and electrons are ejected from it, thus creating an electric current through the wire.

As bold intuition may suggest, the larger the wave is (i.e. the larger the distance between the peak of a crest and the bottom of a trough is, see Figure 1.3, called the *wave amplitude*) the more energy it carries along. In other words, when thinking of electromagnetic waves, intensity and energy are proportional; when thinking of water waves, the taller the wave the greater the energy it carries. To be precise, a typical sinusoidal wave of electromagnetic radiation has an average intensity,

$$\langle I \rangle = \frac{c \cdot \epsilon_0}{2} E_0^2,$$

where $\langle I \rangle$ represents the *average* of the variable I (intensity), c is the speed of light in vacuum, ϵ_0 is a physical constant¹⁴, and E_0 is the maximum electric field strength.

Of course, if we increase the amplitude, an increment in energy will become immediately identifiable; this is true in every aspect of the classical domain. We should then expect a wave to be able to transfer more energy onto an object, a particle for instance, if we increase its intensity. Moreover, if light behaves indeed as wave, and only as a wave, all frequencies would cause electrons to be ejected from the metal plate. If directly exposed to it, electrons should be able to absorb electromagnetic radiation, and transform its energy into *energy of motion*, what we previously referred to as *kinetic energy*. As expected, the amount of energy should be proportional to the intensity of such radiation, and we should be able to detect this by directly measuring the energy of the particles being ejected from the surface of the metal plate.

The intriguing nature of this phenomenon, however, arises when, despite a noticeable increment in the intensity of red light, no increase whatsoever in

¹⁴permittivity of free space

the electric current can be detected. No explanation of this experiment can be provided in terms of classical wave theory, and it was not until 1905 when a satisfactory explanation was achieved, one which relates electromagnetic waves (in the form of light) to a corpuscular nature.

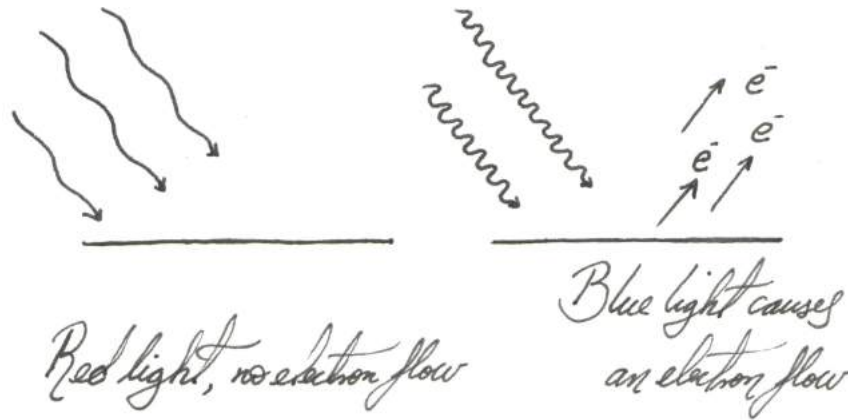


Figure 1.8: The photoelectric effect

As we know from the experiments performed during the 18th century, Young's experiment in particular, light was already, unequivocally, proved to be a wave. Light manifests interference patterns when passed through a two-slit barrier, which makes it qualify as a wave, according to our previous definition. On the other hand, it was when A. Einstein (1889 - 1955) proposed to consider light as a stream of particles rather than as a wave, that the absolute incomprehension of the photoelectric phenomenon came finally to an end.

From what had been observed so far, for any given metal plate that could produce this photoelectric current, light of low frequencies did not produce the desired effect. One could gradually increase the frequency (e.g. going from red to violet), and still no current would be detected. However, from a certain boundary called the *threshold frequency* and further, a clearly detectable flow of electrons was triggered. Moreover, as one increased the light's frequency even further, the kinetic energy with which the electrons were released would also increase, i.e. it became directly proportional to such frequency.

Any proportionality relation can be mathematically expressed by means of a linear function of the corresponding variables. I.e. if x and y are directly proportional, then $y = ax + b$, where a is the proportionality factor and b is an initial value. This *initial value* corresponds to a displacement, at the origin, in the graph of such function. That being the case, the photoelectric effect can be modelled by the following relation (Figure 1.9):

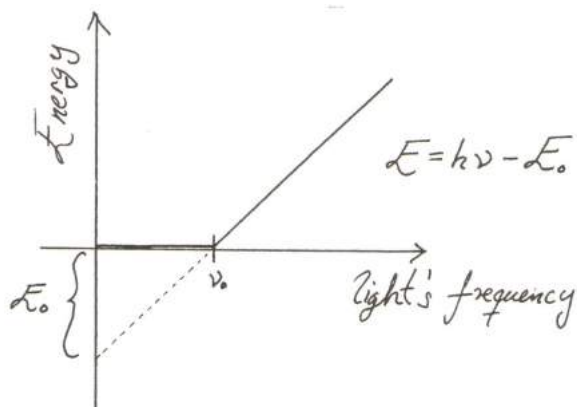


Figure 1.9: The proportionality factor h is known as Planck's constant, as we shall discuss in the next two sections. E_0 corresponds to the initial value; in this case, it is the amount of energy needed for one electron to be released from the metal plate. This value will vary between different metals. ν_0 is the threshold frequency, i.e. the "first" value for light's frequency (colour) for which a current is detectable.

†† If light were to behave as matter, it should be composed of particles. These light-particles came to be known as *photons*, and each photon carries along with it an energy proportional to the light's frequency. Since higher intensities do not produce an increment in the electric current *or* the kinetic energy of the ejected electrons, the amount of energy that each individual electron can absorb is limited to that of a photon, which is itself inherently related to that specific light's frequency. This suggests that light can behave as small, fixed amounts of energy, and not only as a wave with arbitrary energies. Photons are massless and travel through any medium as light beams. In other words, light manifests *both* particle *and* wave-like properties.

†† One of the many conundrums this assertion conveys is that, just as matter can be located in bounded regions of space, one should then be able to tell with a high level of exactness where a photon is. Waves are not something we can locate, they are spread over large regions of space. So, should we be able to locate light in narrowly bounded regions, or should we assume light is spread out all over space as an electromagnetic wave? How would a formal, mathematical description be possible?

§3 Some Necessary Historical Notes

§3.1 Blackbody Radiation

Unless one has infrared or ultraviolet vision, walking into an absolutely dark room means one is (momentarily) completely blind. It does not mean that no electromagnetic radiation is present inside the room, for as we have previously

discussed, the human eye is just sensitive to a small portion of the electromagnetic spectrum. We do have, however, other ways to *detect* radiation. Imagine a room with no visible light, but with an active water heater located in one of the corners. Most people would be able to locate it precisely without the need of visible light. This is due to the emission of a certain invisible radiation commonly known as *heat* (infrared radiation) whenever a certain temperature is reached. If one heats up an object, and moves a hand nearby, it will be noticeable how the intensity of such radiation changes; wherever the heat change increases the most we know *intuitively* is where the object is. The higher the temperature, the easier it is to deduce the location of the object.

If we keep raising the temperature without burning it, this hypothetical object will start to emit a dim red light; above 600° Celsius (about a third of the average temperature of fire when burning butane on a regular gas-stove) it will start glowing, apparently because of the mere fact of *being hot*. By heating it further, up to 2,000° Celsius for example (about the temperature of burning wood), it will produce an easily recognisable yellow glow, and by increasing the temperature further it will radiate intensely with different colours, going progressively from yellow to green, blue, violet and so on. The temperature at which this object starts to emit radiation depends on the material properties of the object, of course, but the phenomenon is quite familiar, easily identifiable and easy to demonstrate, even in the average kitchen.

The fact that we see this glow of any particular colour does not mean this object is radiating *only* this particular wavelength (that associated to yellow light, for instance). When heated, the object absorbs thermal energy and will continue to do so until it reaches a certain temperature; the moment the object has absorbed enough energy, it starts to radiate it out, and *most* of this radiation is concentrated around a certain wavelength (colour), depending on the temperature it reached at this time. As the temperature increases, the intensity peak of this radiation goes from that of reddish hues to the blueish ones of the spectrum.

The graphs shown in Figure 1.10 depict the distribution of radiation of an object that emits energy in the form of electromagnetic radiation due to heat absorption. The x -axis represents the colour, and the graph shows a peak at the specific colour for which this object radiates the most. For higher temperatures, the object radiates highly in the range of shorter wavelengths (blue or violet). Most objects have a tendency to radiate more intensely in the range of wavelengths they are also prone to absorb. Despite the fact that the temperature needed for any specific wavelength to be the peak of the radiation curve is different for every material, the overall properties of this thermal radiation are totally independent of any physical characteristics of the heated object.

A *black body* is therefore a system which absorbs all of the radiation it receives; it is an idealised object, but a cavity with reflecting walls and a pinhole on one side is an excellent approximation. Light comes in through the pinhole, and regardless of the shape of the cavity radiation stays within its walls, thus being “trapped” inside the cavity. Such an object, a black-body, is not per se a system of special interest in physics. It is rather its historic value what gives it

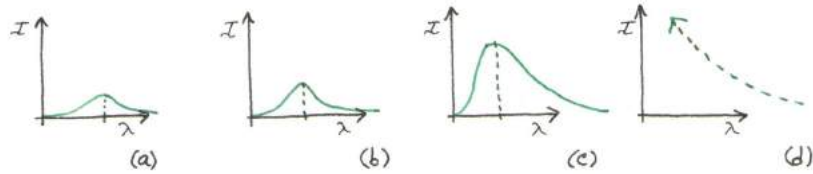


Figure 1.10: Intensity peaks displaced towards violet light as a function of temperature. An object at 300° Celsius (for example) radiates with a peak at a certain colour (wavelength), depicted in (a). Parts (b) and (c) show how this peak is displaced towards the blueish hues of the spectrum as temperature goes up. Part (d) shows the peak's tendency with increasing temperature.

a place in almost every textbook on modern physics nowadays.

The radiation absorbed and then emitted by the cavity in thermal equilibrium does not depend on its colour, shape, form or anything else other than its temperature; for any given temperature T , this black-body will emit radiation of various different wavelengths at their own different intensities, as depicted in the graphs of Figure 1.10.

The electromagnetic radiation at the walls of the cavity must be exactly zero; otherwise, the walls would be infinitely absorbing energy. The radiation that is not absorbed by the interior walls is then reflected, creating an electromagnetic radiation field within the enclosure. That being the case, we can conclude that radiation must exist inside the cavity as *stationary* waves, modelled by sinusoidal curves as the ones depicted in Figure 1.11. A stationary or standing wave has two fixed points, one at each end. The oscillatory motion is therefore restricted to various stable states consisting of nodes and anti-nodes¹⁵. These standing waves have different stable oscillating frequencies, and can be seen in everyday situations like a vibrating guitar string, or waves created by a droplet falling on the surface of a liquid confined within a vessel.

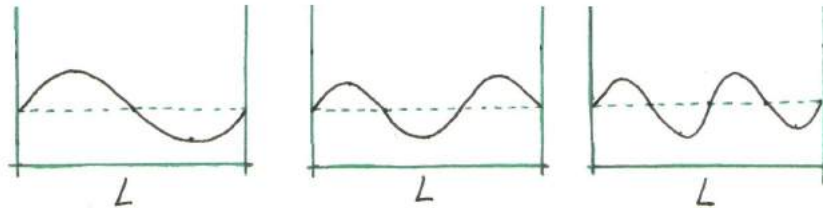


Figure 1.11: Standing waves inside a one-dimensional cavity of length L .

¹⁵Standing waves are discussed in depth in § 6.5 of this chapter.

Statistical physics suggests that the amount of radiant energy per wavelength, i.e. the energy radiated for each wavelength of the spectrum, can be multiplied by the wavelength¹⁶ to obtain the energy corresponding to an average energy and the number of possible standing nodes in that particular range of wavelengths. To be precise, we must consider *all* possible wavelengths, and add up the amount of energy per wavelength times the range of wavelengths we take into consideration, and this sum should be the total radiating energy of the system. Note that this is the actual, finite amount of energy of the black-body.

For the specific case of radiation trapped inside the sealed box, i.e. the black-body, the usual theoretical approach (the only one available before quantum physics) predicts an unbounded tendency for the intensity peak on the side of ultraviolet light. Part (d) of Figure 1.10 shows how the intensity peak's tendency with increasing temperature is unbounded, meaning it can, in principle, reach arbitrarily high values for ultraviolet light. Adding this all up means an infinite amount of energy coming out of the pinhole!

This theoretical prediction is based on the assumption that the energy absorbed and emitted by this black-body can be transferred in any arbitrary amount. The result would be that any such container exposed to a finite amount of energy would radiate back an infinite amount of energy, some of it being lethally energetic, even exceeding the range of X-rays and gamma rays!

Notice that taking all possible wavelengths into consideration means it should be a continuous spectrum of radiation, or so it seemed...

§3.2 What does it Mean to be *Discrete*?

One should always look for the simplest, and preferably most elegant, solution to a problem. On December the 14th, 1900 Max Planck (1858 - 1947), a German physicist, chose the simplest solution to the black body radiation problem. He looked for a mathematical description that actually *fits* the experimental observations. He decided to model the walls inside the cavity as an assembly of linear oscillators that only interact with the radiation field within the enclosure. This simplified the calculations and provided a different value for the average energy (in this case of the oscillators). He then proposed that, instead of considering the whole, continuous spectrum of energy, the energy within the cavity can only be exchanged in multiples of a minimum value, which he called a *quanta*. I.e. the energy inside this container was not assumed to be free to have any arbitrary value, but actually restrained to a set of specific values, all of them multiples of a certain “base” energy that should be proportional to the light's frequency¹⁷, which coincides with Einstein's solution to the photoelectric effect.

†† The idea of considering the existence of certain quantities, like that of matter, light, or electric charge, only in specific amounts instead of a continuous

¹⁶or rather an infinitesimal interval of wavelengths, $d\lambda$ to be precise

¹⁷Frequency is related to wavelength; wavelength is the spatial distance between crests in a wave, frequency is the number of times per second such a wave oscillates through a particular point in space. Frequency is measured in Hertz, meaning the amount of crests passing by for every second of time.

range of values, this “atomisation” of the different quantities in nature, is what we call a *quantisation*. So far we knew that both matter and electric charge are quantised (in atoms and electrically charged particles like electrons); with the discoveries previously discussed, it was made experimentally evident that energy is also quantised. Notice that Planck’s assumption¹⁸ is that energetic interactions between these oscillators and the radiation field *can be modelled* as if they only occurred in fixed amounts, not that the oscillators per se can only possess these fixed amounts. Planck proposed this as a solution to the black body radiation problem, not as a fundamental law of physics; only time and history proved its actual, fundamental quality.

†† By ca. 1910 it was starting to become a well-accepted fact that energy behaves in accordance to Planck’s and Einstein’s quantisation laws: $E = h\nu$, where h is Planck’s constant, E is the energy, and ν is the light’s frequency, and also, any exchange of energy occurs in integer multiples of a minimum amount E_0 or equivalently $E = nh\nu$, where n is an integer¹⁹. Why had we not noticed this in any previous experiment? The smallest macroscopic process involves around 10^{27} *quanta* of energy, thus giving us the false sensation that energy flows continuously in physical processes.

An analogy to understand this concept might come in handy. Imagine a diver about to jump into a pool; he or she is a professional diver, and can choose freely which platform to jump from. Regardless of the amount of heights to be chosen from, every platform has a specific height, and will thus provide the diver with a certain amount of kinetic energy. The diver cannot choose any height he or she feels like choosing, like an intermediate height between two platforms, for example, but would have to pick either one, or the next one. Unlike a ramp, where one can decide in which height to stand (or from which height to jump in this case), platforms constitute a *discrete* set of options to choose from.

†† When our set of possibilities is like a ramp, we say it is *continuous*; when we have to choose one or the next, but nothing in between, we say it is *discrete*. The notes on a piano, for example, are discrete, for they are limited by the keys, whilst the notes on a violin are continuous, for one can press down on the strings at any desired length. Energy, then, was discovered to be always quantised in the microscopic world, i.e. one would always find its values in a discrete set of options, all of them multiples of a first fundamental value of energy called the *ground state*.

¹⁸PLANCK, M. (1900). “Zur Theorie des Gesetzes der Energieverteilung im Normalspectrum”. *Verhandlungen der Deutschen Physikalischen Gesellschaft*.

¹⁹One can also find this as $E = \hbar\omega$, where $\hbar = \frac{h}{2\pi}$, and ω is the light’s angular frequency, a mathematical adjustment to make computations easier. These two expressions are equivalent; they imply that $\omega = 2\pi\nu$. This is how angular frequency and frequency relate.

§4 The Experiment: Electron Diffraction

§4.1 The First Quantum Conjecture

Just from heuristic knowledge of everyday life, we know there is a physical consequence that arises from the interaction with objects that possess mass moving with a certain speed. We are often taught in school that the force an object can exert is equal to the mass times its acceleration, $\vec{F} = m\vec{a}$ in accordance to Newtonian mechanics, but despite what many people could say, this is not at all an intuitive quantity. We experience forces in everyday situations, but we can hardly measure or be sensitive to the acceleration an object is subject to.

What we do know, even before attending any lectures in physics, is that the faster an object is travelling, the greater the effect it can have on us if we decide to step in its way. Obviously, a bullet is barely lethal if it is not travelling at least at a few hundred metres per second. We would rather see a toddler than a bull running towards us, which means mass has also an important effect on material interactions. Just as a snowball rolling down a hillside, objects have a physical property that increases with mass and speed, pointing to the direction of this object's movement. This intuitive property of moving objects is called *momentum*, and its mathematical definition makes its dependence on mass and velocity evident²⁰:

$$\vec{p} = m\vec{v}$$

where \vec{p} is the vector representing momentum, m is the particle's mass, and \vec{v} its velocity vector.

It must be observed that momentum is a physical quantity that transcends these classical definitions. The actual, more general definition of momentum can be more complicated to understand at first glance; it will suffice, for the time being, to keep this as the proper definition. As it will later be seen, the momentum of a photon, for example, exists only due to the fact that it is travelling through space, either as a particle or as electromagnetic radiation; it therefore exists despite the photon being a massless entity.

According to Einstein's *Special Relativity Theory*²¹, the energy associated to any particle is related to its mass and momentum according to the following relation:

$$E^2 = m^2c^4 + c^2p^2$$

where m is the particle's mass, p its momentum, and c is the speed of light in vacuum. Which means that

²⁰Momentum is definitely a more intuitive quantity than force, but this does not mean it was simple and obvious to scientists from the beginning; it took no less than two centuries to formalise these notions into what we call *mechanics*.

²¹Which is out of the scope of this text

$$E^2 = m^2 c^4 + c^2 \cancel{p^2}^0 \implies E = mc^2$$

for particles at rest, and

$$E^2 = \cancel{m^2 c^4}^0 + c^2 p^2 \implies E = cp$$

for massless particles.

†† So from Einstein's and Planck's results on the experiments relating the energy of a photon with the frequency of the electromagnetic radiation, one can conclude that a photon's momentum ($m = 0$) is related to its wavelength λ via the following relations:

$$E = cp$$

&

$$E = h\nu$$

Which implies that

$$cp = h\nu \implies p = \frac{h\nu}{c} \quad \text{and, since} \quad c = \lambda\nu$$

(because speed is measured in metres per second, wavelength is measured in metres, and frequency in units per second)

we get that

$$p = \frac{h\nu}{c\lambda} \implies p = \frac{h}{\lambda}$$

†† Notice how this equation relates a purely oscillatory property, the wavelength λ , and a typically material quantity, the momentum p . This is precisely what was needed, a way to express with mathematical precision how the wave-like and matter-like characteristics of light are related. What was unexpected, though, was the foundational quality of this last relation between a photon's momentum and wavelength via Planck's constant. In fact, what Louis de Broglie (1892 - 1987)²² noticed is the realm of validity of this equation; he identified

²²Louis de Broglie is considered to be one of the founding fathers of quantum physics. He came from an aristocratic French family, and his role in WWI was rather important, for he worked as a communications engineer at the top of the Eiffel tower during the war.

it as a fundamental fact, one that could be generalised to corpuscles of matter. This is, essentially, the first quantum conjecture.

†† Does this mean that every macroscopic body has an associated wavelength? If so, how and where is the associated wave located? Of course, the de Broglie wavelength of macroscopic objects is preposterously small, as can be easily proved by computing the associated wavelength of an everyday object²³, thus exhibiting the frontiers of quantum knowledge. Quantum mechanics works for objects in the atomic range, not elsewhere. This does not mean that the classical description is wrong per se, it is just a macroscopic approximation of the physics in the quantum realm that we happened to discover first because of our macroscopic scale.

So,

$$\left. \begin{array}{l} \bullet E = h\nu \\ \bullet E = cp \end{array} \right\} \text{Then, } \lambda = \frac{h}{p} \text{ for photons.}$$

†† Energy is proportional to frequency via Planck’s constant. The associated “wavelength” of *any* quantum particle, electrons in particular, is related to its momentum via de Broglie’s relations:

$$\lambda_e = \frac{h}{p_e}$$

§4.2 Davisson & Germer

The fact that this relation holds for electrons was proposed in 1924 as de Broglie’s doctoral thesis. No one was sure at the time if such a generalisation could be valid, so the thesis was sent to A. Einstein. The thesis was then accepted, and de Broglie obtained his PhD, but no immediate attempts to test this assumption were made. By 1927, plenty of experiments had already been conducted to confirm both the wave-like and the corpuscular properties of light. Electrons, on the other hand, had been successfully isolated and could be handled in such a way as to perform properly controlled experiments.

At the Bell Telephone Laboratories, in New York, C. J. Davisson (1881 - 1958) and L. H. Germer (1896 - 1971) were studying the reflection of a beam of electrons being shot at a nickel target. An accidental break in the vacuum system created an accumulation of oxide on the nickel surface. This accumulation caused a scattering of the electron beam, which led to the accidental (and

²³Planck’s constant is $h \approx 6.626176 \times 10^{-34}$ Joules · second, which means an object of 1 gram moving at 1 metre per second would have an associated wavelength of about 6×10^{-31} m, which exceeds by far any capacity of measurement. The size of a subatomic particle, like a proton, is roughly 10^{-16} m; this is why quantum mechanics does not work on the macroscopic realm.

by all means interesting) discovery of electron diffraction. As it often happens in science, a discovery was made by mere accident and not by means of the scientific method.

†† Electrons had hit the crystallised oxide, which had worked as a double slit barrier, and were scattered in an indisputable diffraction pattern, with the maxima and minima that any wave would produce. With this, not just light, but electrons were finally proved to be able to manifest both a wave-like and a corpuscular behaviour, thus proving de Broglie’s hypothesis right. Using a crystal as a double slit barrier and the proper experimental arrangement, it is possible to reproduce this results, i.e. the electron interference on a screen. Electrons were always recognised as matter; with this, their wave-like nature was factually exhibited.

§4.3 The Wave Equation

Let us review the results of these experiments involving electron diffraction. After crossing the two-slits barrier, or its crystal equivalent, electrons exhibit an apparent deviation and an alternate pattern of constructive and destructive interference. The screen on the background (a photographic plate, for instance) shows the maxima and minima where electrons “landed” on the screen. One can easily identify a greater concentration of impacts in the central stripe, as the lateral ones appear to get dimmer when being farther away from the centre. The next figure depicts how this electron-impacts-distribution would look like, but most importantly, it shows a graph relating the concentration of such impacts as a function of position. The higher the peaks, the more impacts on the screen are visible. The wave associated to this diffraction pattern is a key starting point for quantum theory.



Figure 1.12: The “wave” that might have caused this interference pattern was perhaps Schrödinger’s starting point. This notion constitutes the basis of the *ensemble* (or statistical) perspective of quantum mechanics.

†† By associating a wave-like description to the experiment, one can describe the *statistical* results obtained, not necessarily attempting to describe the intrinsic nature of a specific electron, but trying instead to encapsulate the observations as a whole. Since the diffraction pattern reflects the statistical incidences of electrons on the screen verbatim, the wave-like description merely shows global characteristics of the system, and *does not necessarily describe the*

behaviour of an electron individually. Whatever the electrons *do* between the barrier and the screen is not a topic of interest to this description of nature. For any purposes of quantum mechanics, electrons could very well act randomly, and this stochastic²⁴ behaviour of individual electrons would not impede a consistent, statistical, collective description. The notion of trajectory is out of the scope of this quantum mechanical description.

There are not abundant sources that can provide historical accounts of how Erwin Schrödinger (1887 - 1961) came up with his ideas in the first place, but in 1926 he published an article with what is now known as *Schrödinger's equation*²⁵. This equation cannot be derived from classical physics. In an article from the 29th volume of *Physics Today* magazine (1976), Felix Bloch (1905 - 1983), a Swiss-American Nobel prize laureate for physics, gives his personal account of the story²⁶:

Once at the end of a colloquium I heard Debye saying something like: "Schrödinger, you are not working right now on very important problems anyway. Why don't you tell us some time about that thesis of de Broglie, which seems to have attracted some attention." So, in one of the next colloquia, Schrödinger gave a beautifully clear account of how de Broglie associated a wave with a particle and how he could obtain the quantization rules of Niel's Bohr and Sommerfeld by demanding that an integer number of waves should be fitted along a stationary orbit. When he had finished, Debye casually remarked that he thought this way of talking was rather childish. As a student of Sommerfeld he had learned that, to deal properly with waves, one had to have a wave equation. It sounded quite trivial and did not seem to make a great impression, but Schrödinger evidently thought a bit more about the idea afterwards.

Just a few weeks later he gave another talk in the colloquium which he started by saying: "My colleague Debye suggested that one should have a wave equation; well, I have found one!" And then he told us essentially what he was about to publish under the title "Quantization as Eigenvalue Problem" as a first paper of a series in the Annalen der Physik.

[...]

there was afterwards a lot of talk among the physicists of Zurich, including even the students, about that mysterious "psi" of Schrodinger. In the summer of 1926, a fine little conference was held there and at the end everyone joined a boat trip to dinner in a restaurant on the lake. As a young "Privatdozent", Erich Hückel worked at that time on what is now well known as the Debye-Huckel theory of strong electrolytes, and on the occasion he incited and helped us to compose some verses, which did not show too much respect for the great professors. As an example, I want to quote the one on Erwin Schrödinger in its original German:

²⁴stochastic = random

²⁵SCHRÖDINGER, E. (1926). "An Undulatory Theory of the Mechanics of Atoms and Molecules" *Physical Review* 28.

²⁶BLOCH, F. (1976). *Heisenberg and the Early Days of Quantum Mechanics*. *Physics Today*, 29 (12), 23-27. doi:10.1063/1.3024633

*“Gar Manches rechnet Erwin schon
Mit seiner Wellenfunktion.
Nur wissen mocht’ man gerne wohl
Was man sich dabei vorstell’n soll.”*

In free translation:

*“Erwin with his psi can do
Calculations quite a few.
But one thing has not been seen:
Just what does psi really mean?”*

Well, the trouble was that Schrödinger did not know it himself. Max Born’s interpretation as probability amplitude came only later and, along with no less a company than Max Planck, Albert Einstein and de Broglie, he remained skeptical about it to the end of his life. Much later, I was once in a seminar where someone drew certain quite extended conclusions from the Schrödinger equation, and Schrödinger expressed his grave doubts that it could be taken that seriously; where upon Gregor Wentzel, who was also there, said to him: “Schrödinger, it is most fortunate that other people believe more in your equation than you do!” Schrödinger thought for a time that a wave packet would represent the actual shape of an electron, but it naturally bothered him that the thing had a tendency to spread out in time as if the electron would gradually get fatter and fatter.

[...]

Schrödinger’s next papers on wave mechanics appeared shortly, one after the other. I did not learn about the matrix formulation of quantum mechanics by Heisenberg, Born and Pascual Jordan until I read that paper of Schrödinger’s in which he showed the two formulations to lead to the same results. It did not take me too long to absorb these new methods, and I wish I could confer to the younger physicists who read this article the marvellous feeling we students experienced at that time in the sudden tremendous widening of our horizon. Since we were not burdened with much previous knowledge, the process was quite painless for us, and we were blissfully unaware of the deep underlying change of fundamental concepts that the more experienced older physicists had to struggle with.

[...]

†† There is [a] [...] remark he once made that I consider even more characteristic. We were on a walk and somehow began to talk about space. I had just read Weyl’s book “Space, Time and Matter,” and under its influence was proud to declare that space was simply the field of linear operations. “Nonsense,” said Heisenberg, “space is blue and birds fly through it.” This may sound naïve, but I knew him well enough by that time to fully understand the rebuke. What he meant was that it was dangerous for a physicist to describe Nature in terms of idealized abstractions too far removed from the evidence of actual observation. In fact, it was just by avoiding this danger in the previous description of atomic phenomena that he was able to arrive at his great creation of quantum mechanics. In celebrating the fiftieth anniversary of this achievement, we are vastly indebted to the men who brought it about: not only for having provided us with

a most powerful tool but also, and even more significant, for a deeper insight into our conception of reality.

From this point on, a whole plethora of experiments were performed both to prove the corpuscular and the wave like nature of electrons. All of them seemed to be decisive, so neither one of the options could be said to be a full description. Meanwhile, and mathematically speaking, classical descriptions of mechanics were failing to provide a proper explanation for the observed behaviour. One should not, however, underestimate the scope of classical mechanics; it was precise enough to help humanity discover planets, galaxies, particles, and electromagnetic spectra, and make a very accurate description of both the microscopic and macroscopic Universe we live in. Keeping the political agendas of the time aside, it even allowed humanity to land on the moon, back when scientific endeavour was a priority.

†† Classical mechanics focuses on the trajectories of particles; it is based on the assumption that, once knowing the position and velocity of a particle, its whole physical history is unambiguously determined, its past and its future are no longer unknown, and classical equations describe the particle's complete existence. Since it is precisely from the recently acquired factual knowledge that Erwin Schrödinger introduced his famous equation, a problem arises when such concrete and tangible notions as localised particles, well-defined velocities, trajectories, etc, are left out from the quantum description of Nature. Is it absolutely statistical or can something actually be said about individual particles? The answer seems to be much more complicated than it seems a priori.

§5 Waves and Particles: One, Both, or Neither

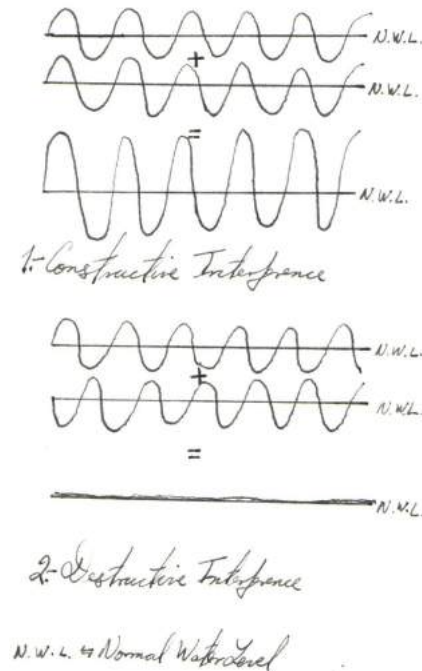


Figure 1.13: Constructive and destructive interference. When two waves are perfectly aligned, they create constructive interference.

Waves consist in travelling disturbances. Sound, for example, propagates through air as a chain of compressions that, once they reach the human eardrum, make it *resonate* in synchrony with the pitch of such sound. Ocean waves are disturbances with transverse displacement with respect to the surface of the water. Radio and television signals are electromagnetic oscillations that periodically change the intensity of an electromagnetic field across the distance between the antenna and your receiver. Normally, these disturbances propagate through a medium, but since the beginning of the 20th century, it was well known to physicists that light needs no medium to propagate. The experiments of Michelson and Morley confirm this hypothesis and unequivocally show that no such medium is required. Any discussion on the subject is worth reading, but it remains out of the scope of this text²⁷.

²⁷Anyone interested in the history of special relativity can read more on the experiments, and find a detailed explanation in the following texts: TIPLER, P. A. (1970) *Foundations of Modern Physics*. Worth Publishers, Inc. Second edition. and BEISER, A. (2003) *Concepts of Modern Physics*. McGraw-Hill. Sixth edition.

‡‡ Although light-waves consist of a *chain* of disturbances, they need no medium to propagate. Matter, on the other hand, *is* something; it does not transfer anything unless it moves. Matter and waves are intrinsically distinct in nature; at least in the macroscopic realm, they are completely different, mutually exclusive phenomena. At the beginning of the 20th century, the physicists involved in the study of the atomic structure noticed that a wave packet, seen from afar, can resemble a particle, a fact that seems quite sensible. However, Louis de Broglie suggested²⁸, in 1924, that matter could be regarded as a wave. Perhaps he postulated this as a mathematical curiosity, as a symmetrical counterpart of seeing waves as matter. Later, Davisson and Germer actually saw an interference pattern whilst shooting electrons through a crystal, which, for experimental purposes, can be considered as the microscopic equivalent of a double slit barrier²⁹, even though electrons had already been proved to be matter. The ensemble as a whole had exhibited wave-like behaviour, but exactly a wave of *what* remained an unanswered question.

‡‡ Once enough experimental observations were gathered, both phenomena were regarded as coming from a common physical manifestation. Most texts on the subject discuss this topics with the term *wave-particle duality*. By using this seemingly inoffensive terminology from the beginning, one may accept that light (or matter) *is* actually both a wave and a particle. Not often does one read any other interpretations, but for this purpose electrons can also be thought of as well-defined particles with well-defined trajectories, but with an intrinsic, associated but separate wave that guides their path and influences their surroundings. If this were the case, then there would be an actual, physical particle, and an actual physical wave, instead of a simultaneous, dual entity. De Broglie favoured this idea, and he spoke about it often, as one can read in his correspondence with Alfred Landau³⁰. He called this a *double solution* to Schrödinger's equation, which considered both an actual, physical wave, as well as the abstract, statistical wave proposed by Schrödinger³¹. As an alternative, electrons can be neither waves nor particles, but something else, something that exceeds our current understanding of Nature, and whose physical manifestations we relate to waves or particles, depending on the experimental situation, but just due to a lack of a better description.

It is noteworthy to state that most advances in quantum theory were done in Europe during WWII and the preceding years. W. Heisenberg's and Schrödinger's contributions were not at all trivial, but suddenly, both Einstein's and de Broglie's perspectives were somehow left aside. The warfare applications of quantum theory were evident, and countries' (like those of Germany and the USA) objectives were focused on the atomic bomb. Quantum physics was on its early stages back then, but a great amount of mathematical development took

²⁸De BROGLIE, L. (1926). "Ondes et mouvements". Solvay Conferences.

²⁹THOMSON, G. P. (1927). "Diffraction of Cathode Rays by a Thin Film" *Nature*. 119 (3007): 890–890.

³⁰DE BROGLIE, L. (1971) *A New Interpretation Concerning the Coexistence of Waves and Particles*. The MIT Press, Cambridge

³¹More on the double solution and the pilot-wave theory is discussed in chapter 3

place, so much that it exceeded by far the actual heuristic understanding of the subject.

Since physicists back then were convinced of the corpuscular nature of electrons, electron diffraction made no sense. The seemingly contradictory observations were part of the blossoming of a new theory, a whole new perspective that we might call *quantum mechanical thinking*. Up to this day no one knows for sure how to comprehend fully these observations, but what we have is a monumental theoretical apparatus that matches every last detail of the experimental results³². However powerful and satisfactory a mathematical tool quantum theory is, the great problem that prevails is that it has no properly satisfying physical interpretation.

§6 So, What Does it Look Like?

§6.1 Quantum Theory

So far we have discussed a few historical facts that led to the development of quantum theory. We stated some of the problems with the description of physical phenomena, and how they were confronted; also, we claimed that Schrödinger's equation was a solution to the so-called *wave-particle duality* problem. We have not yet, however, discussed how quantum theory works, how the wave equation actually solves any of the problems, or even what quantum mechanics takes as experimental facts to describe the physics of the atomic domain in a proper mathematical way. In lieu of an absolutely formal mathematical description, we begin by formulating how quantum mechanics describes its realm of study.

¶¶ Quantum physics is basically the study of the microscopic universe, where classical descriptions are no longer valid due to a set of physical quantities that have no net effect in the macroscopic world, but have large, noticeable effects at the atomic and subatomic level. A quantum system is composed of the portions of the microscopic world one considers for study. Every quantum system has an associated wave-equation³³; this is no experimental fact, it is a starting point of the theoretical “artillery” of quantum physics.

¶¶ This wave equation contains **all** the information needed to describe the quantum system³⁴ fully. It evolves with time, but some of its main characteristics remain unchanged. One usually denotes the wave equation associated to the system of study with the Greek letter Ψ . This wave-function is usually (though not always) a function of time and position:

$$\Psi(\vec{x}, t)$$

³²This is discussed both in § 8 and in Chapter 3

³³A more profound approach to quantum theory can also introduce a *density matrix*. More on this can be found in: DE LA PEÑA, L. (2003). “Introducción a la Mecánica Cuántica.” Fondo de Cultura Económica.

³⁴a *pure* quantum system, to be precise; a mixture of quantum states is again a quantum state, but quantum states that cannot be written as a mixture of other states are called pure quantum states. More can be found at: “Quantum State” at *Wikipedia, The Free Encyclopedia* 2018

where the vector \vec{x} denotes the position in 3-D space, and is downsized to x whenever one deals with one dimensional problems.

So for every point in space, and every moment in time, the wave equation assigns a complex number, which means it is a *function* from the set $Space \times Time$ to the set of complex numbers³⁵. The reasons for this function to be complex are not at all evident. Complex functions allow us to convey more information in very succinct forms; also, they are often easier to handle. A few other properties of complex functions should be enough justification, but for the moment, it can simply be seen as a mathematical imposition.

¶¶ Despite the fact that this wave equation has no physical interpretation, it is considered to contain all the relevant quantum information we can ask about the system. This does not mean, as we will clarify further in the text, that one can “ask” the wave equation for *all* that we wish to know about such system, for the knowledge of some of the variables may influence our knowledge of the rest of them. This is one of the most intriguing facts of quantum theory. Let not this hinder our enthusiasm about quantum mechanics just yet; the wave equation can answer a lot of our questions about the system, and it has continued to do so for most of the past one hundred years.

§6.2 The *Size* of the Wave Function: A Probability Distribution

Probability is a very general branch of mathematics, but its role in quantum mechanics is quintessential, since most of the experimental procedures one can do with quantum systems consist in obtaining average measurements related to specific physical quantities. So, in order to understand better how quantum theory is structured, let us digress for a moment and discuss some relevant facts about probability³⁶.

Whenever one has a set of statistical data, there is only a range of possible outcomes for any given experiment; that is to say, the *set of all possible outcomes* should be a well defined set of parameters. For example, by tossing a die one expects to obtain a whole number between one and six, with a certain well-defined probability, but one would never expect such a die to fall in $\frac{\pi}{2}$, of course, since it is not even in the range of possibilities. The set of possible outcomes is then $\{1,2,3,4,5,6\}$, which we shall label Ω . In this case, and assuming the die is perfectly homogeneous and even, each number has a strict $\frac{1}{6}$ probability of appearing in every toss.

A probability distribution for a discrete variable is called a *probability mass function*; it is a function of a discrete, random variable x , labelled $p(x)$ that assigns a probability, a number between 0 and 1, to every member of the set of discrete possible outcomes. In the previous example, such a distribution would assign $\frac{1}{6}$ to each number in the set $\Omega = \{1,2,3,4,5,6\}$. The graph of

³⁵It can also be a function from the set of momentum parameters and time, to the set of complex numbers. As a first approach, and for simplicity, one can understand it as it is here presented.

³⁶More on this and other topics on probability are discussed in Chapter 2

this particular distribution would be a simple set of horizontally aligned points. Adding consecutive values of the function $p(x)$ corresponds to the probability that this random variable x lies between these specific values. Adding every value of the function along all points of the set Ω must be equal to 1, since the probability of getting *some* value of Ω is 100%.

A probability distribution for a continuous variable is called a *probability density function* (or pdf for short), and is usually labelled $\rho(x)$ (not to be confused with $p(x)$), where x is the random variable of interest. It provides the amount of probability per given range of values of x . In the case of age, for instance, the question *what is the probability of a random person being 26 years, eight months, four days, two hours, one minute, six seconds... old?* is absolutely meaningless, since the probability of someone having any specific age is strictly zero. One must allow a range of possible values, however small we want it to be.

Let Δx be the range of age we happen to be interested in. If $\rho(x)$ is the probability density, measured in “probability per range of age”, the value of $\rho(x) \cdot \Delta x$ is an actual probability. If we want to refine our measurement, we can narrow down the range Δx we are interested in. Loosely speaking, we label an *infinitesimally* small range “ dx .” For the value of $\rho(x)dx$ to make sense, using this infinitesimally small number, we can add a range of values and obtain the probability that the random variable x lies precisely in the interval between a and b .

$$P(a, b) = \sum_{j=a}^b p(x_j) \quad \implies \quad P(a, b) = \lim_{\Delta x \rightarrow 0} \sum_{j=a}^b \rho(x_j) \cdot \Delta x = \int_a^b \rho(x) dx$$

For a discrete variable

For a continuous variable

where the symbol $\int dx$ represents a sum for a continuous variable.

This can be seen intuitively if one thinks of very, very small intervals Δx that partition the interval $[a, b]$. These Δx are the bases of very thin but tall rectangles of height $\rho(x)$; as one adds up the area of these rectangles, each of them represented by the product $\rho(x) \times \Delta x$, i.e. the base times the height of the rectangle, the resulting number corresponds to the area of a region underneath the curve $\rho(x)$. As this partition is refined, i.e. as we consider smaller and smaller intervals Δx , this resulting number approaches the exact value of the area underneath the curve. It is also the probability of the variable x having a value between a and b , i.e. along the whole interval.

†† Probability distributions ought to be representative of the whole collection of experimental data; it should not be understood to represent exactly which the particular outcome will be. It is impossible to predict any particular outcome. In fact, any statistical information comes from a set of data; it reveals the *tendency* of a given variable, provided a random event. If the testing conditions

change, then that same variable x is not “prepared” in the same state, and the information might seem biased. E.g. The average grade on a group of students is a value that represents the whole set of grades; it does not mean it is the most repeated value or even that an actual student got this particular grade. It is usually helpful to resort to averages instead of actual values; average values can be essential, and sometimes even more useful than precise data.

Perhaps an example can be more effective to illustrate this point. If one wishes to study the dynamics of a passenger train, for instance, and one needs to calculate how much fuel it takes to drive it from one city into the next one, it would be useless only to consider the weight of the train. It needs fuel to move, and the heavier it gets, the more fuel it requires. Since it will also be carrying passengers, staff, and their luggage, one has to take a lot more mass into account than that of the train alone. Every trip would be different, with a different amount of passengers, baggage, etc, and even the weight of any given passenger can be different on different days. The only way to estimate the amount of fuel needed is to consider what the *average* trip looks like, i.e. how many people travel between these two cities in average, how many pieces of luggage they carry, and so on. Notice how this data is even more significant than any given measurement of a certain specific trip, however precise it were.

Average quantities are very important when dealing with large sets of data. Thermodynamics, for example, is a branch of physics that relies heavily on statistical data and average values of physical quantities. It would be practically impossible and even useless to have specific data about mass and velocity for every molecule in a cubic metre of gas, especially since it contains an average of 10^{25} molecules, but the average kinetic energy of its molecules allows us to predict macroscopic variables such as temperature or pressure. Many problems at the early stages of quantum mechanics were dealt with mathematical methods that come from statistical thermodynamics.

†† A typical wave function in quantum mechanics contains, specifically, the statistical information needed to obtain the average values for different physical quantities. The average position, momentum, kinetic, and rotational energy, etc, are examples of exactly the kind of information the wave equation has. One “asks” the wave function for the average value of a particular physical quantity by performing a measurement. Figure 1.14 represents the probability density function for a particle located in a one dimensional line. Any region underneath the curve $\rho(x)$, bounded by two values a and b , represents the probability of finding the particle within that interval in space; these values are always bound to be a number between 0 and 1.

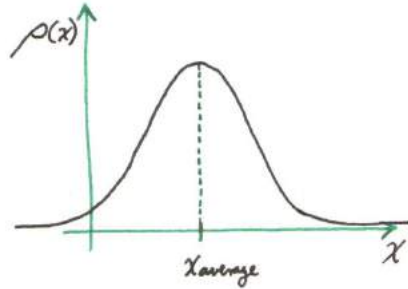


Figure 1.14: This particular probability density function for a particle in space is often known as a *normal* or *Gaussian* distribution. For every range dx in space, the density function $\rho(x)$ is a curve that shows the probability density in that particular region. Here, x_{average} is the average value (also referred to as *expectation value*) for x , which does not mean one will actually find the particle there, just that it is more likely according to the distribution. The equation for such a distribution is the following: $\rho(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-x_{\text{average}})^2}{2\sigma^2}}$, where x_{average} is naturally the average position, and σ is related to how “far”, in average, every element is from the average value.

The entire space is represented by the x -axis, i.e. values of x ranging from $-\infty$ to ∞ . Since the particle must be located at *some* point in space, by adding up the individual probabilities of finding the particle at each point in space, we end up having the 100% certainty (thus the number 1) of locating the particle *somewhere*, i.e.

$$P(-\infty, \infty) = \int_{-\infty}^{\infty} \rho(x) dx = 1$$

†† As was discussed in §4.3, classical physics is based on the assumption that, once a particle’s initial position and velocity are determined, its whole physical history, i.e. its whole past and future are theoretically determined. Classical equations are deterministic; in contrast, a quantum mechanical description of a physical system contains all the information needed to make a measurement and obtain a precise result, but given its statistical nature, any event at the quantum level *can be considered* random, and the theory provides the probability of such an event to take place.

†† As far as quantum mechanics is concerned, anything that happened before the measurement was made can be considered as non-existent, and the mere act of performing a measurement *destroys* any previous information that the wave function had provided, depending on the quantity one measures. This is not a trivial fact; on the contrary, it is a key feature of quantum physics, one that defines the core difference between the macroscopic realm and the

quantum world. Does this mean that Nature itself is random, and that the particle's physical reality was in fact undetermined before the measurement was made? No, not necessarily, it just means the theory per se does not contain this information; it is not part of its mathematical description.

†† A typical wave function is depicted in Figure 1.15; it is defined in every point in space, and as its name suggests, its graph is usually related to a kind of wave-like motion. One cannot emphasise enough that the wave function has **no actual physical meaning**, despite the fact that it contains all the physical information of the system. By squaring the wave function, however, we obtain what we can call the *size*³⁷ of the wave function. The graph of this new function does have a physical meaning, it is the probability density function for the given quantum system, a particle for example, at any given point in space³⁸. To be more precise,

$$|\Psi(x)|^2 = \rho(x)$$

This interpretation of the norm of the wave function is due to Max Born (1882 - 1970), a German mathematician and physicist also considered a key figure in the development of quantum theory. The “size” of the wave function (or any function) refers to the values the function takes, but ignoring their orientation. So, for example, if an arbitrary function represents the relative position of an object with respect to a certain point in space, the values this function takes can be positive, meaning the object lies “after” the reference point, or negative, meaning the object lies “before” the reference point; the size of this function would represent the absolute distance between the object and the reference point.

§6.3 Measurements

Whenever a measurement is made, the physical system of interest is disturbed, at least ever so slightly. This is true in classical physics, and it is true in quantum mechanics. Take, for instance, a classical measurement of temperature in a macroscopic system. We can suppose the system is originally at a temperature T_0 ; when the system and the thermometer are placed in contact, a certain amount of heat is transferred to the mercury (or alcohol) of the thermometer to make it work, thus altering the system's original temperature. However, the main difference between both cases, classical or quantum, is the following:

(a) The theoretical model of classical mechanics considers every object to be equivalent to a point in space (its centre of mass). Everything has a determined, well-defined position and momentum, and one can measure them both. The precision of such measurements depends only on our equipment, but it can be, in principle, arbitrarily narrow. Once the object's momentum and position are

³⁷Also called the *norm*. The *size* or *norm* of a function is actually just $|\Psi|$, and here we have written $|\Psi|^2$; this temporary nomenclature can be useful for clarity.

³⁸For more about Born's perspective see BORN, M., (1969). *Atomic Physics*. Blackie, Eighth Edition.

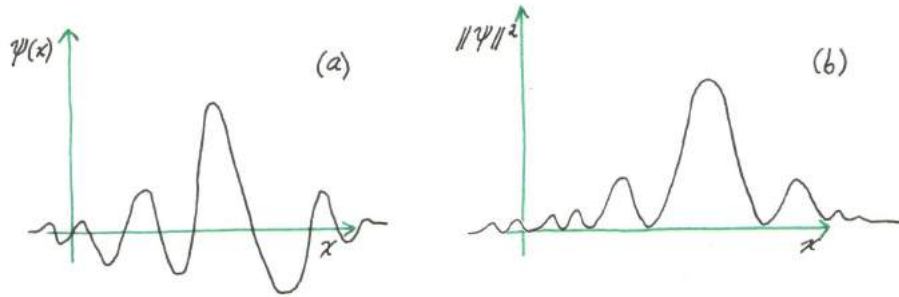


Figure 1.15: Part (a) shows a typical (real) wave function, whilst (b) represents its *size* squared. The x axis represents space. The size of the wave function squared is a probability distribution, meaning that wherever this function takes higher values, is where the particle is *more likely* to be found. As one moves farther away from the system (far from the laboratory, for instance), the size of the wave function gets closer, and closer to zero, which is expected, since it is related to the probability of finding the particle there. It is obviously unlikely to find the particle very far away from the actual experiment.

known, Newton's equation provides everything needed to deduce the particle's entire existence. The more forces one takes into consideration, the wider the range of validity of the resulting equations. Trajectories through space and time are the core of Newtonian mechanics, and somehow we have all learnt to think in terms of this physical model³⁹.

†† (b) Quantum mechanics models each system with the aid of a wave function, and the notion of *trajectory* is inevitably lost; it is not properly defined, and not even considered in the theoretical model. If they exist or not at the microscopic level is irrelevant for quantum theory, and certain interpretations of experimental evidence often *suggest* they do not. Throughout the years it has made it ever more difficult to define such a notion, since quantum particles disobey some elementary rules of classical mechanics. Measuring a particle's position means affecting the wave function in such a way that it seems to "force" the particle to be where we found it to be.

†† The wave function evolves in time, it changes, but whenever we perform a measurement and find a particle to be at some location x_0 , the size of the wave function approaches zero elsewhere, and creates a large probability peak around x_0 , as if the particle was materialised precisely there as a consequence of our measurement. Why there and not somewhere else? That is an unanswered question of quantum physics. Up to now, we consider this behaviour to be simply stochastic. As time goes by, the wave function re-stabilises, and recovers its form, spread out across a large region of space.

†† The only fundamental property that remains unchanged throughout this

³⁹despite the fact that, in most cases, the resulting system of classical equations cannot be solved analytically

whole process is the fact that the particle *has* to be somewhere, and thus the area under its probability density curve should keep adding up to 1. The form of the wave function changes, but the area underneath it remains unchanged. See Fig. 1.16

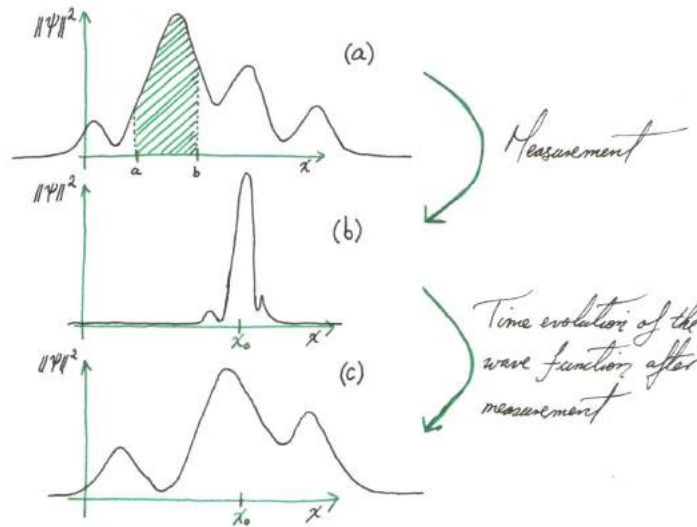


Figure 1.16: The probability density for a typical wave function defined on the x axis. Part (a) shows the meaning of $P(a, b) = \int_a^b \rho(x) dx$; it is a real positive number between 0 and 1, i.e. the probability of finding the particle in the interval $[a, b]$. Part (b) shows what happens to the size of the wave function when a measurement is made; here, the particle was found to be in a very narrow region around the value x_0 . Part (c) shows how, with time, the wave function and its size evolve to a more stable state, spread out over a larger region in space, without losing the fundamental property $P(-\infty, \infty) = \int_{-\infty}^{\infty} \rho(x) dx = 1$, the particle is still *somewhere*.

§6.4 The Typical Problem (I)

Most of what one learns whilst studying quantum physics, or almost any other subject in physics, is how to solve problems. Problems in quantum mechanics can be very hard to solve, but they often involve finding the associated wave function for a given physical system. The purpose of this section is not to teach the algebraic processes that lead to a solution of any particular problem, but to provide an overview of how problems in quantum mechanics are usually dealt with. Consider, for example, the following situation:

There is a quantum particle, like an electron, confined to a region of space by means of two impenetrable walls. The electron is trapped in a one dimensional space, meaning it can only move in a straight line, i.e. left or right, but neither backwards or forwards, nor up or down, like in an extremely thin tubular structure. These “walls” can be in fact a pair of transverse cuts around the region where the electron is confined in such tubular structure, two cuts that do not allow the electron to flow to other regions of space. A voltage difference applied with a pair of batteries⁴⁰ forces the electron to stay within the boundaries of this inescapable “box.” These specific conditions that describe the regions of space where the quantum particle is “allowed” or not are called *boundary conditions*; once defined, one can ask how this simple quantum mechanical system will evolve, according to the laws of quantum mechanics⁴¹.

The first thing to do is to translate this physical problem into a mathematical one; this is often the most complicated part of the task. After that, one can use any mathematical tools one considers to be useful and find a mathematical solution. Then, one reinterprets the solution again in physical terms. Finding a solution without the use of mathematical formalism would be nearly impossible; this is why one often requires the aid of precise mathematical formulation, it allows us to focus on the relevant variables, and to find unequivocally true relations between them. To solve this kind of problems, we first get rid of any extra information, and retain only whatever we consider indispensable.

For simplicity, and since this quantum system is considered as one-dimensional, we represent *space* by a segment of the x axis, and the boundaries are taken to be absolutely impenetrable. The first task at hand is to find the associated wave equation, which should evolve according to Schrödinger’s equation. By taking both the Schrödinger equation and the appropriate boundary conditions, one can successfully present this problem in the language of formal mathematics.

⁴⁰A much more detailed explanation of this experiment and many other similar ones can be found in the first chapters of DE LA PEÑA, L. (2003). “Introducción a la Mecánica Cuántica.” Fondo de Cultura Económica.

⁴¹It should be noted that a real experimental arrangement would be much more complicated; to begin with, everything should be done in a vacuum chamber, etc.

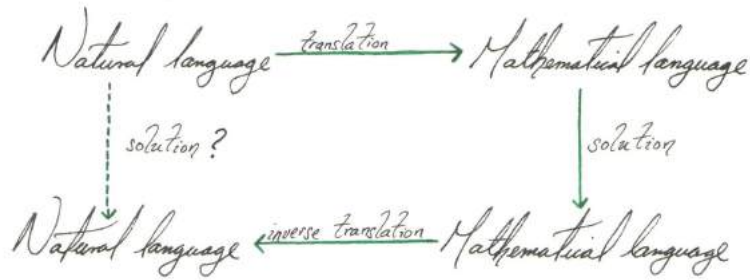


Figure 1.17: The process of problem solving in science. Both “paths” commute, but solving the problem with the aid of mathematics is much simpler, precise, and concise.

$$-\frac{\hbar^2}{2m} \nabla^2 \psi(\vec{x}, t) + U(\vec{x}, t) \psi(\vec{x}, t) = i\hbar \frac{\partial}{\partial t} \psi(\vec{x}, t)$$

Schrödinger's equation

Figure 1.18: Schrödinger's equation. The description of mechanical systems at a subatomic level

This means we are looking for a function $\Psi(\vec{x}, t)$ that for every position in space and every moment in time assigns a complex number z , very much like the one depicted in Figure 1.15. This function should be such that whenever twice differentiated, and multiplied by the factor $\frac{-\hbar^2}{2m}$, i.e.

$$\frac{-\hbar^2}{2m} \nabla^2 \Psi(\vec{x}, t)$$

and then multiplied by the spatial constrains, represented by a function $V(\vec{x})$ that depends on the points in space \vec{x} (that means it describes the allowed regions as a function of position \vec{x})⁴², i.e.

$$V(\vec{x})\Psi(\vec{x}, t)$$

when added, will be precisely equal to its time evolution, represented by the function's time derivative $\frac{\partial \Psi}{\partial t}$, and finally scaled by the factor $i\hbar$, i.e.

$$i\hbar \frac{\partial \Psi(\vec{x}, t)}{\partial t}$$

Again, the goal of this text is not to teach a step by step solution to this problem, or even a formal mathematical derivation, much less algebra or calculus. The goal, however, is to demystify some of the aspects of quantum physics, and present them as a starting point for a much deeper understanding.

So, if the symbolic representation $\frac{\partial \Psi}{\partial t}$ does not mean anything to you, just read it in your mind as the *time evolution of the wave equation* Ψ ; it tells us how the function changes as time goes by. The “nabla” operator, ∇ , is related to how the wave function Ψ changes from place to place. The fact that it is squared, meaning it should be performed twice, is related to the particle's kinetic energy $E_k = \frac{1}{2}mv^2$. A broader explanation can be found in *Appendix 3: Arriving at the Schrödinger Equation*. Finally, as it was stated earlier, the function $V(\vec{x})$ describes the regions in space where the particle is allowed, and the “effort” involved in reaching such regions. It should suffice to say that in order to understand how these mathematical operations work, one must first familiarise oneself with the notions they represent.⁴³

⁴²A more intuitive approach to this kind of functions is found on section §7.5 of this chapter, *Quantum Tunneling*

⁴³This is the way we truly learn any subject. Take language as an example; one does not learn English (or any other language), by reading thick, extensive books on grammar, structure, etc first, and then going out to the world and trying to apply it; we grow acquainted with it, we get involved in culture, we listen to the language in the most diverse situations throughout our lives, we hear thousands of conversations we do not understand, we misspell and mispronounce words we are not even sure of their meaning sometimes, we even use grammatical structures we do not fully understand, but eventually we learn the language. It is good advice, if you want to learn physics, to get acquainted with its language, its structures, while also reading the formal books about it.

§6.5 Digression: Harmonics of the *Mikrokosmos*

A violin, a guitar, a cello, a piano; they all work under the same principle. Taut strings whose ends are tightly fixed to the instrument are made to vibrate, either by striking them or rubbing them. These oscillations displace molecules of air in a periodic motion; the train of air compressions travels through space and into our ears; eardrums are then displaced accordingly, moving back and forth, matching the frequency of the oscillations of the string in the instrument. Two separate instruments can begin to match a sounding frequency just by standing in the way of a travelling sound-wave.

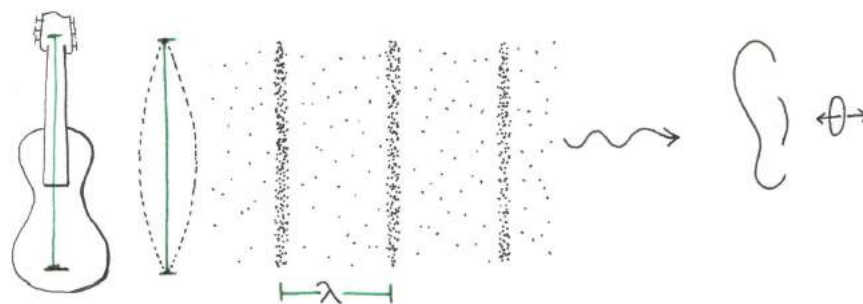


Figure 1.19: When striking a guitar string, the oscillations push the air molecules every other time, creating a chain of air compressions that travel through the air and reach the eardrum. The wavelength λ is related to the distance between compressions. The eardrum resonates matching the sounding frequency of the guitar string.

Every time one presses down on a string, the distance between fixed ends is changed. The string, however, can oscillate stably in different ways without anyone pressing down at some point throughout its length. The different ways a string can oscillate without changing the distance between the two fixed ends are called *stationary modes*. This means that the transverse wave motion of the string can survive without losing or gaining energy, for it is a stable travelling wave that reflects on both ends, but leaves fixed points called *nodes*. The resulting wave pattern is also referred to as *standing waves*.

Figure 1.20 depicts a taut string (of a guitar, for example), where only the ends are fixed; it shows the different ways it can vibrate, i.e. the stationary modes of vibration. The intermediate points of equilibrium are the nodes, whilst the points of maximum displacement are called *anti-nodes*. The first harmonic, also called a fundamental in music theory, corresponds to striking a string without holding it tight at any point. The second harmonic is reached by merely touching the string exactly at the middle point and making it vibrate; this will create two first-harmonic-like waves, each on one side. Since the new lengths are different, the pitch will be different.

As one carries on producing harmonics, the pitch goes up, as one can see in

the Fig. 1.20. This is very easy to recreate on a guitar or a violin. Pressing down on any other point along the string will produce a travelling wave that is not stationary, meaning that the fixed ends work as reflection points, and the wave dissipates rapidly.

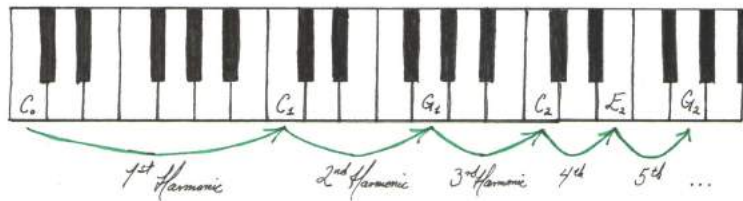
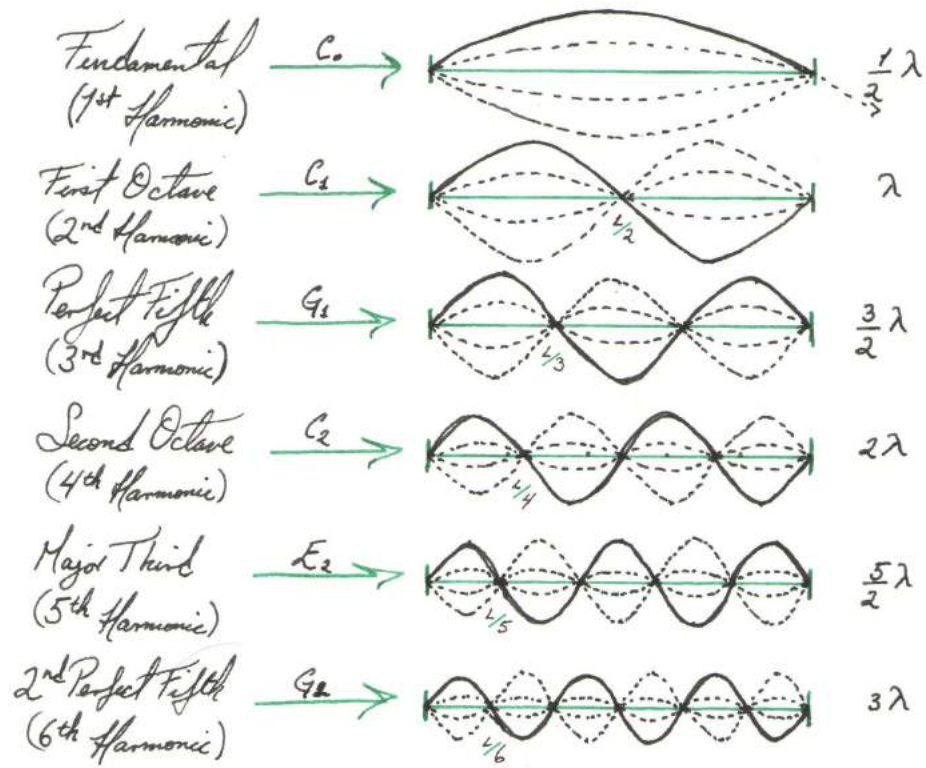


Figure 1.20: Harmonics in music, and the harmonic progression on the piano keys.

Stationary modes	$L = k\lambda$	$k \in \{\frac{1}{2}, 1, \frac{3}{2}, \dots\}$
λ_1	$L = \frac{1}{2}\lambda_1 \implies$	$\lambda_1 = 2L$
λ_2	$L = \lambda_2 \implies$	$\lambda_2 = L$
λ_3	$L = \frac{3}{2}\lambda_3 \implies$	$\lambda_3 = \frac{2}{3}L$
λ_4	$L = 2\lambda_4 \implies$	$\lambda_4 = \frac{1}{2}L$
...

Table 1.1: Stationary modes and the relations between the length L and harmonic wavelengths λ

§6.6 The Typical Problem (II): The Solution to the Schrödinger Equation

Going back to the problem of a quantum particle trapped in a one dimensional box of impenetrable walls, we want to find a suitable wave function that describes this particular problem. We stated earlier that such a solution should satisfy both the boundary conditions, and the Schrödinger equation, providing all the physical information of our system that is available at the atomic and subatomic level. Let us not forget that the size of this wave function squared will allow us to predict where the quantum particle is likely to be, for it is the probability density function for positions \vec{x} in space.

The solutions to the Schrödinger equation are **exactly** those of Figure 1.20, i.e. the musical harmonics. This means that a quantum particle, trapped in a linear segment (like a copper wire) with impenetrable boundaries provided by the voltage difference of a pair of batteries, is subject to a wave function that increases harmonically in half-segments of a fundamental wavelength λ . Let us take a moment to look at the wave motions on Figure 1.20; when seen as the harmonic scale of a string tuned to C, the length of the interval represents the length of the string. The first harmonic (or fundamental) represents this pitch, and corresponds to half a wavelength, since the whole wavelength is measured from crest to crest, or trough to trough. (see Figure 1.3). The rest of the harmonics correspond to the following ways the string can vibrate without any extra energy, i.e. in a stable, stationary mode.

When seen as a solution to the problem of the quantum particle though, the length of the interval represents the length of the one dimensional box, i.e. the wire segment where the particle is trapped. The oscillation modes correspond to valid wave functions, or in other words, valid solutions to the Schrödinger equation. The first solution corresponds to a particle in an energetic *ground state*. The next solutions are increasing levels of energy, each having an associated wave function with half a wavelength more than the previous energy state, i.e.

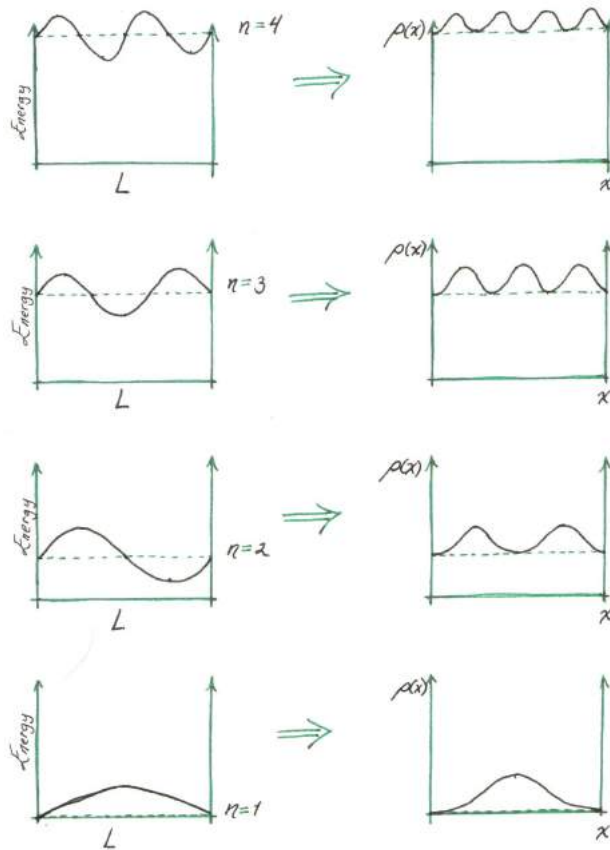


Figure 1.21: Graphs on the left depict the wave function solutions for the first four levels of energy. Their corresponding probability distributions on the right exhibit where the quantum particle is likely to be located. Notice how every node implies a point where the particle will *certainly* not be found.

¶¶ In summary, we want to find a wave function that properly describes the quantum system of interest; we impose the physical constraints proper of this system, e.g. boundary conditions, and we ask that this wave function obeys the fundamental property $P(-\infty, \infty) = \int_{-\infty}^{\infty} \rho(x) dx = 1$, i.e. that its size squared is in fact a probability density function (and we say it is a *normalised* function). Finally, we need this wave function to satisfy Schrödinger's equation, so we plug it in, and by means of mathematical tools we find the wave function associated to our quantum system.

So, quantum mechanics reveals some deep interconnections in science; it brings a whole new perspective into our physical understanding of the Universe, and it compels us to develop a new intuition that breaks from the paradigm of

classical physics, but why is it so intriguing? How is it that after a hundred years of its birth it still beguiles the physicists community, and why do we find it so provocative?

§7 Why is it so Intriguing?

†† Throughout the decades, people have been fascinated by the conceptual implications of quantum mechanics, mostly because they defy very basic notions of our physical, heuristic experience with the world around us. Some other notions confronted by this new paradigm are not precisely evident for someone who is not a physicist, someone who has an “outsider’s” understanding of physics, for they involve the debunking of certain concepts of classical mechanics that require previous knowledge, and the development of a refined intuition in the first place.

†† E.g. to find it intriguing that a quantum particle’s angular momentum is quantised, one first has to understand formally and intuitively what angular momentum is. Quantum physics does have an interesting set of topics, of unanswered questions that provoke a deep, thoughtful rethinking of our most basic concepts, but these topics are now so extensively popular that they have often been misinterpreted. We now discuss some of these fascinating topics.

§7.1 Superposition

Adding two wave functions results in another wave function; if both of these satisfy the adequate boundary conditions, and the area underneath their sizes’ functions adds up to 1, then the resulting sum will also satisfy this set of conditions, making it a new valid wave function for the quantum system of interest. In other words, *two valid solutions to the Schrödinger equation of a quantum system can be added, and the resulting wave function will also be a valid solution*. This result can be extended to any number of solutions, so any arbitrary number of valid wave functions can be summed to obtain a new valid solution, even an infinite number.

This is called the *superposition principle*. A quantum system is mathematically described by a wave function; it is said to be in a *quantum state* that is either a pure state or a mixture of different quantum states. Two or more states can be added, despite their physical meaning, and the new associated wave function will still be valid. So, for example, the wave function representing a particle’s rotation being *strictly* to the left can be added to the wave function representing the same particle’s rotation being *strictly* to the right to create the quantum state of a single particle that is in a simultaneous state of left and right, according to the mathematical description of reality at a quantum level.

†† Superposition is mathematically consistent, but leaving this aside for a moment, one has to stop and consider the interpretational issues this induces. The sum of these two different wave functions should be a valid representation of the state of a quantum system, despite the coexistence of two possible, seemingly contradictory notions. A similar macroscopic analogue would be, for example,

the wave function associated to the toss of a coin that is still in mid-air to be the sum of the two possible states, the coin landing heads-side up, **and** the coin landing tails-side up, simultaneously contributing to a unique, stable state of existence. Notice that the state of the coin is not heads *or* tails, but heads *and* tails.

There are at least two ways to understand superposition:

(a) A quantum particle's state is *described*, mathematically, by a superposition of quantum states that comes from a previous statistical knowledge of an ensemble of similar particles. The physical variables of the particle are in fact determined, and an act of measurement will result in our knowledge of the state of such variables, with probabilities determined by the size of the wave function. (Realist perspective)

(b) A quantum particle *is* in fact in a quantum superposition of states. Its physical variables are undetermined before any measurements are performed. These will randomly take form in one and only one of the possible outcomes, once the measurement is done. The probability of this particular outcome is given by the size of the wave function. (Orthodox perspective)

The problem can be translated into a macroscopic paradox⁴⁴. The idea of a macroscopic system attached to a quantum particle was presented by E. Schrödinger in his famous cat experiment, where a quantum system exists in a stable superposition of seemingly opposite states, along with a Geiger radiation detector ready to activate a poison trap inside a box, depending on the physical manifestation of the state of the system, and a cat trapped inside this box. If the system does actually exist as a superposition of quantum states, then the Geiger counter also exists in a superposition of activation/non-activation, and thus the cat remains in a stable dead/alive state of existence.

The mathematically congruent fact of adding two valid solutions into a single new valid solution brings a new, problematic interpretation into the theory. The verisimilitude and possibility of a macroscopic superposition of states remains an unsolved puzzle in physics. Option (a) discards this possibility at once.

§7.2 Quantisation

As we discussed in a previous section, the solutions to the Schrödinger equation for a particle confined in a one dimensional box are the series of harmonic modes. This series does not stop at the sixth harmonic, but carries on infinitely into smaller and smaller wavelengths, each one being more energetic. This means that not every possible wavelength represents an acceptable solution, but just those whose half multiples fit exactly in the length L of the box (recall the relation $L = n \cdot (\frac{1}{2}\lambda)$, where n is a natural number, i.e. $n \in \{0, 1, 2, 3, \dots\}$), thus the idea of a *quantisation* of energy, and the name of *quantum physics*. Recall

⁴⁴It is a paradox if one does not accept the notion of quantum superposition for macroscopic systems, which seems quite reasonable.

that being quantised⁴⁵ means being discrete, bound to be exactly one or a next one, but not any arbitrary quantity in between. For this particular case, energy being quantised means that a quantum system is only allowed to “occupy” a certain state from a set of infinite, discrete energy states.

The energy of this system is related to the stationary modes via the following equation: $E_n = n^2(\frac{\pi^2 \hbar^2}{2mL^2})$, where \hbar is Planck’s constant⁴⁶, m is the mass of the quantum particle, L is the size of the one dimensional box, and n is again a natural number. This is in perfect resonance with M. Planck’s and A. Einstein’s postulate of the quantisation of energy for light.

Allowed values for energy

$$E_0 = 0$$

$$E_1 = \frac{\pi^2 \hbar^2}{2mL^2}$$

$$E_2 = 4 \frac{\pi^2 \hbar^2}{2mL^2}$$

$$E_3 = 9 \frac{\pi^2 \hbar^2}{2mL^2}$$

$$E_4 = 16 \frac{\pi^2 \hbar^2}{2mL^2}$$

...

⁴⁵Not to be confused with *quantified*, which means to have an associated quantity, as in *Love is something that cannot be quantified*.

⁴⁶As defined before, \hbar is Planck’s original constant divided by 2π

This is certainly a major difference with macroscopic physics; it implies that every physical process involves the exchange of energy *only* in a multiple of the minimum energy amount E_1 , called *quanta*. The fact that energy seems as a continuous variable in the macroscopic realm is only due to the extraordinarily large amount of these small quanta involved in everyday physics, whose discrete nature is impossible to identify at the human scale.

†† I.e. for the electron trapped inside the “box,” the allowed values of energy start with E_0 , which means there is no particle at all, but the fact that there is a minimum amount, E_1 , means that trying to trap the electron in an ever smaller box implies an enormous increase in its kinetic energy, due to the large oscillations of its wave function.

§7.3 “Collapse” of the Wave Function

Quantum mechanical descriptions of physical phenomena come from the observation of collective phenomena. Isolating an electron is an extremely labourious and complex task to accomplish, and isolating photons was nearly unachievable until quite recently, so most of the experiments of particle physics are done with large ensembles of particles. Atoms are so small, that the tiniest speck of dust contains trillions and trillions of them, which means that the sharpest experimental precision accounts only for the measurement of (very good) averages of physical variables. This is in no detriment to a successful physical theory; as we discussed earlier, averages can even be more significant than specific data, and this statistical nature is not necessarily a flaw of quantum theory. Statistical physics was in vogue during the last half of the 19th century due to the vast success of thermodynamics, and many of its mathematical techniques were useful during the development of quantum theory.

Physics is usually known for its predictive nature; unlike physics, probability describes no real physical experimentation, but abstract mathematical scenarios. E.g. everyone knows that the odds of tossing a fair coin and finding it to have landed heads-up is exactly 50%, no more, and no less. Does this mean that it comes up heads every other time? No, it definitively does not. Does it mean that if we flip it ten times in a row, five of them will come up heads? No, it does not, as everyday experience easily proves. Then, what does it mean, really, that the probability of heads coming up in the first toss are 50%? Can we predict the outcome of this experiment by knowing this statistical information? Again, the answer is *no*.

Probability tells us what the *overall tendency* of an experiment is, but only after thorough repetition. In the case of the coin, it means that tossing the coin any certain number of times will approximately produce half of the outcomes as *heads*, and half of the outcomes as *tails*. It means that *if* we were flip the coin “exactly” an infinite amount of times, only then would be find exactly half of the outcomes being *heads*. Since the notion of tossing a coin “an infinite amount of times” makes no sense, one can simply describe the formal mathematics, and predict the tendency of the experiment. One cannot predict the outcome of any particular toss, but merely speculate about the whole set of experimental data.

Quantum systems are analysed with the aid of statistically deduced equations, and it should not be surprising that they match any particular set of collective data, but during the years physicists have used these equations as if they were a correct and complete description of *individual* quantum particles, and this has constantly suggested to be an appropriate assumption. Nowadays, Schrödinger's equation is mostly interpreted as a full description of any individual quantum particle, from the range of protons and neutrons to electrons and quarks, and this comes with several interpretational problems.

If a quantum system exists in a stable superposition of states, its associated wave function will be composed of a (finite or infinite) sum of different wave equations, each one representing a different quantum state, i.e.

$$\Psi(\vec{x}, t) = a_0\psi_0(\vec{x}, t) + a_1\psi_1(\vec{x}, t) + a_2\psi_2(\vec{x}, t) + \dots + a_k\psi_k(\vec{x}, t) + \dots = \sum_{j=1}^{\infty} a_j\psi_j(\vec{x}, t)$$

where any two functions $\psi_k(\vec{x}, t)$ and $\psi_j(\vec{x}, t)$ may represent seemingly contradictory states (as being dead *and* being alive, in Schrödinger's cat example), and the coefficients a_k represent the probability amplitude of the system to be in the state $\psi_k(\vec{x}, t)$.

¶¶ We know from experience that neither cats are ever in a dead/alive superposition, nor are objects in two different locations at once, but the associated wave functions of quantum particles seem to imply they are. However, when we perform a measurement, as we discussed in a previous section, the wave function seems to concentrate highly around a precise state x_0 , which is the state where we find our quantum system to be. This is usually referred to as the *collapse of the wave function*, because the whole wave equation, which is spread out over large regions of space, “collapses” into a single state. I.e. the quantum superposition collapses into one of its constituent states. The probability density function creates a large peak around the specific value x_0 , where we found the particle. This can be interpreted as if the particle was *not* there before the measurement, but everywhere else, and was rather *created* there because of the measurement (Orthodox view). This interpretational complexity derived from an orthodox perspective has troubled the physicists community for decades, and no convincing answer seems to satisfy everyone completely.

§7.4 Heisenberg's Inequalities: “Uncertainty”

Heisenberg's inequalities are undoubtedly one of the most controversial, thought-provoking topics in quantum theory. The original German word, “*Unschärferelation*,” is perhaps better translated as *un-sharpness relation*, which is not exactly the same as uncertainty. We now discuss three different approaches to understanding what these “uncertainty” relations mean.

Physical Experiment: An Intuitive Approach

We are able to see objects around us because light bounces off their surfaces. Besides the particle behaviour of light, its wave-like character as electromagnetic oscillations is useful to understand plenty of the natural phenomena of our everyday lives. Light travels around us as this electromagnetic flux⁴⁷, and we find our way through space by locating objects via the light they reflect. Some animals, like bats, are mostly blind; they find their way (significantly better than we do, even through complete darkness) by producing a noise (imperceptible to the human ear) and locating objects around them, i.e. their brain interprets the way disturbances in the air bounce off the objects of their surroundings, and helps them build a perfectly sharp image of the place around them. Bats are very good listeners, but based on our definition of seeing, they are almost absolutely blind. But, *are* they? It is clear that some of our definitions hold merely to the realm of human understanding, and we want to go much further than that.

When we enter the subatomic world, observing (measuring) a quantum system seems actually to alter the outcome of the experiments, and even more counter-intuitively, the possibilities of making a precise measurement at this scale become very limited. In §6.2 we said that a measurement alters any information of a quantum system prior to the measurement. The confidence with which we can measure certain variables can be narrowed down to obtain a better, more accurate result. The means by which we narrow down this margin of error can affect the accuracy to measure another variable.

In 1927, Werner Heisenberg published an article⁴⁸ about the “*Un-sharpness*” *relations*, his latest results at that time. These can be interpreted as to establish an impossibility, not just experimentally but also theoretically, of measuring both the position and the momentum (or velocity) of a particle simultaneously. In any other world-view or scientific perspective it is fundamental to know the attributes of the objects of study, and never is it questioned if one can (either practically or theoretically) know them or not; if one considers an orthodox interpretation of quantum mechanics, this limitation lies in the formal mathematical structure of the theory.

However hard or impossible it may be in practice to know the location of something with absolute precision and exactness, no one ever doubts that an object (e.g. a particle) *is* really somewhere. One thing is not knowing its location, and something very different is to recognise that the idea of physically *being* somewhere is per se nonsensical. Part of the paradigm shift in quantum mechanics is to accept that many of the common and elemen-

⁴⁷Any course on Electromagnetism can be useful to understand more about the nature of light. For a very precise and clear explanation see: GRIFFITHS, D. J., (1999). *Introduction to Electrodynamics*. Prentice Hall, Third Edition.

⁴⁸HEISENBERG, W. (1927). “Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik”. *Zeitschrift für Physik*.

gate waves

tary notions of the scientific world are meaningless in the subatomic domain. Any two variables whose relation is governed by this principle, are commonly said to have an uncertainty relation, e.g. position and momentum, the energy of a particle at a certain event and the time this event took place, etc. Let us try to understand this fundamental characteristic of nature that seems so reminiscent of Heraclitus' ⁴⁹ fragment, $\Phi\upsilon\sigma\iota\varsigma\ \kappa\rho\upsilon\pi\tau\epsilon\sigma\theta\alpha\iota\ \phi\iota\lambda\epsilon\iota$ (*Nature loves to hide itself*).

To grasp the full extent of Heisenberg's discovery we need to rethink our very notion of seeing. What does it mean, for example, to *see* an electron? What if we were so small that the concept of *seeing* turned out to be preposterous. Macroscopically speaking, we see because photons, particles of light, hit the objects around us, bounce off of them, travel into our eyes, and hit our retinæ. Photons are so small, they can be considered to have length zero, not close to zero, but strictly *zero*. They normally have no influence on objects whilst hitting them and being reflected (if someone crashes into you and bounces off, you will certainly be pushed away as well, but not if a speck of dust impacts against you).

Electrons, however, are so small that they do “feel” the effects of being observed, i.e. of photons crashing into them. They are *deflected* from their normal paths. In order to locate an electron by observing it, we need to perform some measurement, and eventually use light to see it. As regards electrons, light can behave both as a particle and as a wave, so let us take it as a wave for a moment. If we wished to locate the electron with the best of precisions, we would have to use light with a very short wavelength; that way we can trap the particle inside the wave with great precision and exactness. The shorter the wavelength, the more precise our measurement will be. On the other hand, if we wished to know its velocity, we would want to alter its path and energy the least we can.

Evidently, waves with greater wavelength carry less energy than those with shorter wavelength. Think of a baby with floaters bouncing up and down in the middle of a pool. If he bounces slowly and with the least up-down displacement, a bug standing in the water far away from him will hardly notice its presence. If the baby bounces up and down harshly and very rapidly, the bug far away will feel a lot more and will probably be very much affected. So, by using large-wavelength light, we alter the electron the least, and our precision to measure velocity increases. Analogously, by using short-wavelength light, we alter the electron the most, but our precision to determine its position increases. We can now almost clearly see how we sacrifice precision to determine one of the physical variables whilst determining the other and vice versa (See Figure 1.22).

‡ In the words of Max Born⁵⁰, one of the founders of the theory of quantum mechanics, “*The theory speaks of the orbit and the velocity of an electron round the nucleus, without regard to the consideration that we cannot determine the position of the electron in the atom at all, without immediately breaking up the*

⁴⁹HERACLITUS. B123-DK

⁵⁰BORN, M., (1969). *Atomic Physics*. Blackie, Eighth Edition. The original version was published in 1935.

whole atom. In fact, in order to define its position with any exactness within the atom (whose diameter is of the order of magnitude of a few Angström units), we must observe the atom with light of definitely smaller wave-length than this, i.e. we must irradiate it with extremely hard X-rays or with γ -rays; in that case, however, the [...] recoil of the electron is so great that its connection with the atom is immediately severed..."



Figure 1.22: Trying to visualise uncertainty.

Fourier and Conjugate Variables: A More Formal Approach

†† Given a periodical signal, e.g. a musical note, we can plot the air pressure differences at a fixed point in space as a function $f(t)$ of time t . If it were electromagnetic radiation, or any other oscillatory system, the mathematical description would be exactly the same. Take, for instance, a musical note with a frequency ν , measured in Hz, meaning the air pressure differences occur at a rate of ν times per second⁵¹. Given this function $f(t)$, there is another function $\hat{f}(\nu)$ called its *Fourier transform* that provides the composing frequency, frequencies, or range of frequencies. All sound is composed of a combination of frequencies, so this Fourier transform would provide accurate information of these constituent frequencies.

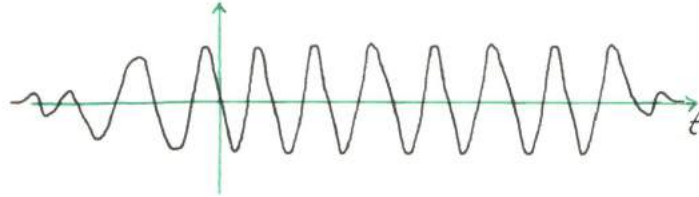


Figure 1.23: A musical note that starts sounding at some time t and is sustained over a long period of time.

This almost pure note would be decomposed into a very narrow range of frequencies. Noise, i.e. impurities in a note's signal, is translated into a broader peak on the function's Fourier transform's ($\hat{f}(\nu)$) graph. A pure note would constitute a single, infinitely narrow peak in its Fourier transform's graph; a chord, the superposition of various different notes, would be decomposed into a series of peaks, located at the constituent frequencies in $\hat{f}(\nu)$. A sheet music is precisely the Fourier transform of music, the notes on the staff being a depiction of the chord's constituent frequencies.

⁵¹E.g. a guitar plays the note A as a vibration of a string at a rate of 440 times per second.

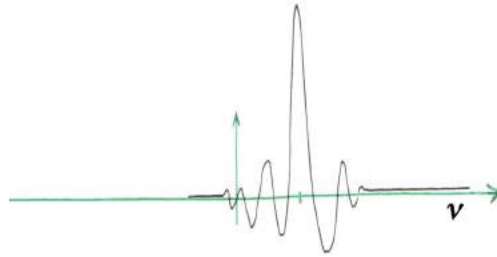


Figure 1.24: Its fourier transform $\hat{f}(\nu)$ has a large peak around the frequency of the musical note, and is close to zero elsewhere.

The Fourier transform can only provide sensible information **if** the musical note is sustained over a large period of time. Naturally, the shorter the note, the more it seems like a click or a hand-clap, and any attempt to decompose it into a series of composing frequencies becomes meaningless.

†† In this case, time and frequency are said to have an uncertainty relation, since a well-defined frequency means the note cannot be located precisely in time; analogously, a sound that can be located precisely in time cannot have a well defined frequency, and its Fourier transform will be spread out across the frequency domain. In quantum mechanics, position and momentum are a pair of such variables, meaning one is the other's Fourier transform, thus the uncertainty involved in their measurements. Mathematically, the Fourier transform is considered a change of coordinate system (e.g. from time to frequency and vice-versa).

Operators and Measurements:

The action of performing a measurement in position is usually denoted with the action \hat{X} ; likewise, performing a measurement in momentum is denoted by the action \hat{P} . Since the act of measuring one of these alters the measurement of the other, performing both actions in different order results in a different outcome. This can be better expressed by means of their commutation rules, i.e.

$$\hat{X}\hat{P} - \hat{P}\hat{X} \neq 0 \iff \hat{X}\hat{P} \neq \hat{P}\hat{X}$$

The action of measuring first position, and then momentum produces a different outcome than the action of measuring first momentum, and then position. In quantum mechanics, *conjugate variables* are defined as pairs of physical observables whose actions (mathematically known as *operators*), e.g. \hat{X} and \hat{P} , do not commute.

†† To give this a more statistical interpretation, we can say that for all the elements of an ensemble of quantum particles to have the same, fixed momentum, all particles should be uniformly distributed throughout the whole space. This does not permit the determination of any particle's trajectory, but only statistic properties of their movement as a whole. This would not necessarily mean that such trajectories do not exist, just that they are not considered in the mathematical formalisation of quantum theory.

If Δx represents how close, in average, every member of the ensemble is to the average position $\langle x \rangle$, whilst Δp represents how close, in average, every particle's momentum is to the average momentum $\langle p \rangle$, then Heisenberg's inequalities can be represented as follows:

$$\Delta x \Delta p \geq \frac{\hbar}{2}$$

Notice how an increment in precision for one variable implies a proportional decrease in precision for the conjugate variable.

§7.5 Quantum Tunneling

If one throws a ball on a flat surface that turns into a small hill, the ball comes back after running uphill for a brief moment; it moves at a certain speed, and slows down as it goes up. After it has reached a certain height, all its kinetic energy is lost, and it comes to a full stop. For the briefest moment, it stops completely, and then comes back to its starting point. How high the ball can go depends on how much kinetic energy we confer upon it; the greater the energy, the higher it goes.

Kinetic energy is related to an object's velocity by the following relation⁵²,

$$E_k = \frac{1}{2}mv^2$$

where m is the mass of the object, and v its velocity. Notice how a higher velocity is translated into a much larger kinetic energy, due to the fact that v is squared.

Potential energy, denoted as $V(\vec{x})$ in a previous section (also $V(x)$ if the problem is one dimensional), defines the regions of space which are harder for an object, the ball for instance, to reach. The hill in the previous example would be one of such places, since reaching the top of this small hill requires a larger amount of kinetic energy. This energy due to motion was transferred to the ball by an external source, and it was not necessarily kinetic energy before being transferred. The idea that energy can be accumulated, and then transformed into a different kind of energy, but never changing a total amount, denoted E_{Total} , is the core of the *law of conservation of energy*. This principle is a central piece in theoretical physics.

⁵²This equation can be derived using calculus; if the reader is not familiarised with calculus, this equation can be taken as a definition.

Energy is conserved by constantly changing its manifestations. For example, nuclear energy is released from Helium atoms in the Sun's core. From there, it is radiated as electromagnetic radiation, and travels through space as light for about eight and a half minutes before reaching the Earth's surface. Plants absorb this radiation, and capture its luminous energy, storing it in molecules called glucose. Animals feed on plants and profit from this energy, saving some of it as fats or muscle tissue. Humans feed on plants and animals, using oxygen to burn out the energy that was saved in meat as sugar or fats by the process of breathing. This energy is released and used by humans to move objects, walk, think, read, etc. Plants do the exact opposite chemical process by using the energy from the sun to transform carbon dioxide and water into food.

Gathering energy and saving it in sugar molecules or any other means is very useful for living beings. Saved energy can be transformed into movement (kinetic energy), heat (thermal energy), sound (an acoustic form of mechanical energy), electricity (electric energy), etc. The fact that it is stored means that it can be transformed into *any* kind of manifested energy, hence the name *potential*.

Imagine a graph like the one in Figure 1.25, where every place in a portion of space has its correspondent height marked by the graph $h(x)$; the two peaks can represent two small hills seen from one side. Place a ball at the point A , and let it roll downhill. It will reach the next peak, go past it, and fly off, perhaps out of sight; i.e. it has enough energy to go past the boundary, and reach the other side. Placing the ball at point B , however, will result in an oscillating, harmonic motion inside the valley between the two peaks. This means that the ball did not have enough energy to cross the barrier into the next region to the right.

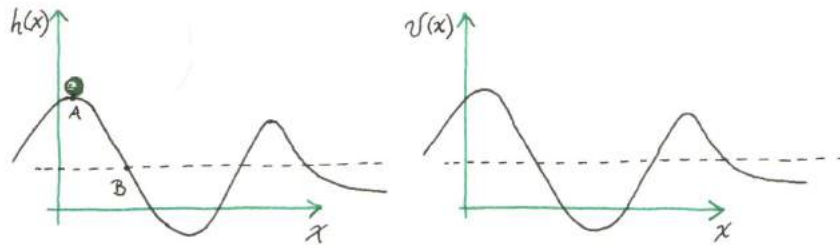


Figure 1.25: Potential wells.

If one can understand the dynamics of the ball in the graph $h(x)$, one can just as easily understand the graph of a potential function $V(x)$ like the one on the right hand side of Figure 1.25. I.e. the dynamics of a particle in a one dimensional space x , subject to a potential $V(x)$, would be that of a ball moving through the heights marked by $h(x)$.

So, in general, a potential function $V(x)$ shows the energetic restrictions for an object moving along the one dimensional space x , and it should be read as we did for the ball and the two peaks, i.e. higher regions are harder to reach, and

thus more energy is needed for an object (a particle for example) to be there. If an object does not have enough energy to cross a barrier, it will stay on one side, either swinging back and forth in the valley (sometimes called a *potential well*), or simply sitting at the lowest point.

†† As any regular particle would do, a quantum particle that is launched from the left against a potential barrier will be reflected if it does not have the right amount of energy to surpass it. The quantum particle, as opposed to a classical particle, is described by a wave function, and not by a well-defined trajectory; its associated wave function will be partially reflected, but also partially transmitted, however small this transmission is. This phenomenon is known as *quantum tunneling*, for it seems as if the particle had “tunneled” its way out of the barrier. It is very counter-intuitive, but also very useful. Its practical applications vary from electronic microscopy to high precision laboratory tools.

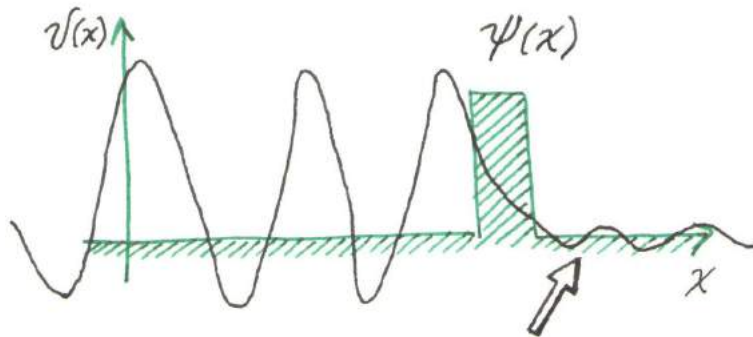


Figure 1.26: The wave function of this particle decreases severely after the barrier, but is not *strictly* zero, which means the particle can, in principle, be found there. It is very unlikely, but it is possible; in classical physics, any particle with this amount of energy will definitely, with 100% certainty, not be found there.

§7.6 EPR

The following discussion is about the phenomenon known as *quantum entanglement*, one of the most controversial phenomena in today’s physics. It enables quantum-computing, and most of the problematic consequences it gives rise to are still an open problem in physics. As we know from Einstein’s famous equation, we can turn energy into matter and vice versa. When matter is created out of energy, both a particle and an anti-particle are created. This process is very much like minting; when one makes a coin, one cuts a round piece of material and, inevitably, ends up with a counter-piece, like the “negative” image of the coin, exactly as round, but *opposite*.

Suppose we create, out of energy, a matter/antimatter pair (e.g. an electron and a positron). These two particles possess an intrinsic, unchanging value for their rotational energy which we call “*spin*,” and every time one of them has an associated value for spin, the other one will have the opposite value. In other words, if the electron has a clockwise rotational energy, then the positron will have an anti-clockwise one.

It is useful and didactic to think of *spin* as an amount of energy associated to an intrinsic rotation of particles. Spin “up” is a good way to label a clockwise rotation, and spin “down” a good way to label an anti-clockwise rotation. It is not something considered by quantum theory, or even remotely measurable, if a quantum particle does actually rotate around its own axis or not; this is a mere approximation to understand a much more intricate property of quantum particles.

Whenever a physical quantity is conserved, there is an associated number that remains constant throughout any physical processes. E.g. the total energy of a system is conserved, and thus the sum of its potential and kinetic energy remains as a constant number throughout any process the system may undergo. That is to say,

$$E_{Total} = E_{Kinetic} + E_{Potential} = constant$$

i.e. a fixed value

†† For a pair electron/positron, this conserved amount is the total spin of the system. If one has spin up, we can associate it a number 1; consequently, the other one has spin down, and we associate the number -1 . This way one could say that the system as a whole always adds up to zero⁵³. Quantum systems like this one are intrinsically connected, so their states are **not** independent, and the whole system is described by the same wave function. This means that whatever happens to one of the particles will have an effect on the other one. Measurements, for example, affect the entire system, and there is a “collapse” of the conjoint wave function. When this happens, the two particles are said to be *entangled*.

Let us clarify this concept by means of a simpler example. We describe the state of a system with the following notation⁵⁴: $|\psi\rangle$. So, for instance, tossing two different coins can result in the following outcomes:

$$\begin{array}{cc} |h\rangle |h\rangle & |h\rangle |t\rangle \\ |t\rangle |h\rangle & |t\rangle |t\rangle \end{array}$$

⁵³A particle’s spin, its “amount of rotational energy,” is *always* quantised, so no value for rotational energy can fall between any two multiples of the accepted ones $\frac{h}{2}$, but for purposes of this example we just consider spin to be 1 and -1 .

⁵⁴The state of the system should be read as “ket psi.” Kets represent states, in this case the label for such state is *psi*.

where $|h\rangle$ represents a coin falling heads-up, and $|t\rangle$ represents a coin falling tails-up. The first ket represents the outcome of the first coin, and the next one the outcome of the second coin. These are not entangled states, for they exist independently of one another.

The cat inside the box with a poison trap, however, cannot have this same arbitrary set of states, because

$|\text{Cat}_{alive}\rangle |\text{Poison}_{inact}\rangle$ is possible, but $|\text{Cat}_{alive}\rangle |\text{Poison}_{act}\rangle$ is impossible;
 $|\text{Cat}_{dead}\rangle |\text{Poison}_{inact}\rangle$ is impossible, but $|\text{Cat}_{dead}\rangle |\text{Poison}_{act}\rangle$ is possible.

Here, of course $|\text{Cat}_{alive}\rangle |\text{Poison}_{inact}\rangle$ means the cat is alive whilst the poison was not activated, $|\text{Cat}_{alive}\rangle |\text{Poison}_{act}\rangle$ means both the cat is alive and the poison was activated, and so forth.

So the pair electron/positron is an entangled system, and can be denoted as⁵⁵

$$|\psi\rangle = |\uparrow\rangle_e |\downarrow\rangle_p + |\downarrow\rangle_e |\uparrow\rangle_p$$

since it exists in a superposition of both states before we perform any measurements.

Recall that the wave function of a quantum system contains all the physical information of the system. This wave function is spread out over space, but concentrates highly whenever we perform a measurement. Position, momentum, spin, polarisation, these are all physical quantities that involve a measurement, and some of them relate via the Heisenberg inequalities.

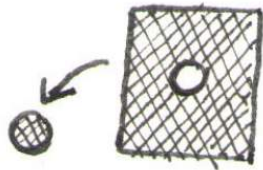


Figure 1.27: Minting results in a coin/anti-coin pair. Creating matter inevitably creates anti-matter

As we discussed before, performing a measurement on any of these quantum particles will result in an alteration of the wave function describing the entire system. These measurements are said to be *correlated*, and it means that if someone measures the electron to have a spin value of 1, the spin of the positron will be immediately determined to be -1 , and the superposition of both states will “collapse” into this well-defined state of existence.

To summarise, consider this pair of matter/antimatter, created out of energy somewhere in the galaxy. They are entangled, and they share the exact same characteristics (electric charge, spin, mass, etc) but in an opposite way⁵⁶. They do so in such a fashion that everything that occurs to the former will be “*felt*” by the latter, as in a physical manifestation of symmetry. Let us say, for instance, that after their creation, the positron travels through the galaxy all the way into a laboratory, and the electron travels to

⁵⁵Of course, no normalisation factors were taken into account, only for the sake of a clear exposition.

⁵⁶except for mass, mass is the same for both since it is a real, positive quantity

the other side of the galaxy, and into another laboratory. Whatever observable, physical quantity one can measure, the other laboratory can do as well. If one measures, for example, its spin to be ‘up’, the other could do the experiment and *always* find it to be ‘down’.

Suppose that once the particles are far enough, one laboratory actually measures the electron’s spin; that would produce an immediate “collapse” of the wave function, causing the immediate response of the positron. This violates the Special Relativity principle that nothing can travel faster than the speed of light, since the information of an experiment being done on one particle was transferred immediately to its entangled counterpart at the other end of the galaxy. Moreover, the first laboratory could also measure its particle’s position with great precision⁵⁷, and the other one, in turn, could measure its particle’s speed. Since both particles left the same place in space with equal speed, but just opposite directions, knowing one particle’s speed implies knowing the other one’s. People from both laboratories could travel all the way through the galaxy to meet each other and discuss the physics of the experiment; they can exchange results and just switch signs. By knowing the opposite values of their results they can infer both particles’ velocity *and* position, which were incompatible variables according to the uncertainty relations. Heisenberg’s relations result from statistical analysis, but they are mostly interpreted as if the accuracy of a measurement of momentum implies a corresponding inaccuracy in the measurement of position, meaning the knowledge of both quantities violates a theoretical impossibility.

This experiment was first proposed by A. Einstein, Boris Podolsky, and Nathan Rosen in 1935, and it is now known as the EPR Paradox⁵⁸.

Solutions to the Paradox

(·) There exists some *spooky action at distance* in between the two particles by which nature will “keep us” from knowing both quantities at the same time. I.e. there will be something to restrain us from successfully performing one of the measurements.

(··) There is some sort of communication between the two particles that instantaneously makes one aware that its partner has been “observed,” and thus telling the other one into which state it should *collapse*. Since both laboratories are located at different ends of the galaxy, this violates Einstein’s postulate, which implies that either Special Relativity is wrong, or quantum mechanics is missing the consideration of a set of physical variables, and is therefore incomplete.

(···) Particles carry a kind of “DNA” called *local hidden variables* that allows them to know the nature of their counterparts before any measurement is per-

⁵⁷We could consider any pair of variables that hold an uncertainty relation. Spin works perfectly, and the real experiment is quite different and more profound. It is here modified and simplified for didactic reasons.

⁵⁸EINSTEIN, A., & PODOLSKY, B., & ROSEN, N. (1935). “Can Quantum-Mechanical Description of Physical Reality be Considered Complete?” *Physical Review* 47

formed upon them, and even know beforehand the state in which they should be found when *observed*. This would imply that, despite the chaotic nature of our Universe, these experiments were bound to happen since the beginning of time, and the outcomes were already a pre-established fact for Nature. Another possibility is that all particles are aware of each other, and thus carry an enormous amount of information with them or, even more interestingly, all electrons in the Universe are one and only, that simply manifests itself differently as in different “quantum states.”

(...) The assumption that a quantum experiment can remain as an entangled system even through large scale, relativistic distances is perhaps too audacious, and even an irresponsible extrapolation. One has to consider that quantum mechanics deals with the physics at the scale of Planck’s constant ($10^{-35}m$), whilst Special Relativity focuses on scales of the order of 3×10^8 metres. In that case, a more realistic task would be to try to delimit the actual boundaries of quantum physics.

There was an attempt during the decade of 1960 to solve these weird consequences, explained in terms of a mathematical inequality⁵⁹, which basically states that EPR does not quite hold for our Universe. The possible incompleteness of quantum theory has been tried to be both demonstrated and debunked. Up to now it is still an open question in physics. It is interesting to note that Newtonian mechanics lasted for more than 300 years; it helped humanity understand an enormous range of scales. It would not be surprising if it took quantum mechanics another 200 years of refinement before we find a better theory.

“The dividing line between the wave and particle nature of matter and radiation is the moment ‘now.’ As this moment steadily advances through time it coagulates a wavy future into a particle past... Everything in the future is a wave; everything in the past is a particle.”

W. L. Bragg⁶⁰

§8 Interpretations

There are essentially two interpretations of quantum theory: the *statistic* or *ensemble* interpretation, and the orthodox, or *Copenhagen* interpretation. Both were briefly mentioned in §7.1 without any further explanation. It is worth mentioning that the Copenhagen interpretation was dominant throughout the 1930’s and 1940’s, and has prevailed throughout the decades. Most of today’s physics is done without any reference to a particular interpretation; quantum mechanics has been very useful to predict phenomena that was impossible to

⁵⁹John Bell’s theorem. BELL, J. (1964). “On the Einstein Podolsky Rosen Paradox”. *Physics 1*

⁶⁰William Bragg. Attributed. BEISER, A. (2003) *Concepts of Modern Physics*. McGraw-Hill. Sixth edition.

describe or understand in terms of classical physics, and the technological advances achieved by its correct predictions are undeniably great in number. These successes have gradually left the problem of finding an acceptable interpretation aside, and physicists nowadays mostly consider it to be irrelevant.

§8.1 The Ensemble Interpretation

The statistical interpretation of quantum mechanics is considered to be a *realistic* approach to the physical consequences of the theory. It can be summarised in the following principles:

- The quantum state $|\psi\rangle$ of a system is a mathematical description that applies only to an ensemble of *identically prepared* quantum systems; it does not represent the state of any individual element of the physical system.
- Quantum mechanics is a statistical theory; it is not necessarily a complete description of physical reality.
- Properties like *trajectories* are not part of a quantum description of physical reality, but this does not mean they do not exist.

The ensemble interpretation requires the fewest assumptions; it need the least “extra” explanations of the physical phenomena.

§8.2 The Copenhagen Interpretation

The Copenhagen interpretation of quantum mechanics can be summarised in the following principles:

- The position, momentum, etc of a quantum particle are, in fact, non-existent before any measurement is done. These variables are not unknown, nor are they out of the scope of the mathematical description, but de facto undetermined.
- The measurement process “creates” the physical reality of the quantum particle; Nature randomly selects the outcome of the measurement from a well-established set of options called the *spectrum* of the operator associated with the measurement.
- The quantum state $|\psi\rangle$ is the superposition of realities that “collapses” into that which we see, the *result* of the measurements. It is the interaction of the “observer” with the quantum system that causes the wave function to “collapse” into one of its constituent states. It does so with greater incidence in those values corresponding to the states in which it is most probable to find the system.

The orthodox interpretation rejects the objective reality of the quantum world. “*Reality is in the observation, not in the electron.*”⁶¹

In the double slit experiment, for example, electrons are waves that go through one slit, the other, both, and none at the same time. Once measured, electrons “become” particles again, and go back to forming the regular two stripes pattern. This is perhaps a good example of the reasons this interpretation is so popular amongst physicists.

§9 A Somewhat Satisfying Justification

¶ Mathematics has always been the language to convey the a posteriori meaning of scientific explorations, and mathematical descriptions often encompass the core of rational thought, of that which we consider to be true and consistent. Mathematical models of natural phenomena usually come from a great cumuli of well-understood knowledge. Before the 20th century, physics had used the mathematical apparatus as an apex of conceptual and factual comprehension. With quantum mechanics it was the other way around, and this particular aspect made it sombre from the beginning⁶². Quantum theories arose from an abstract, unfamiliar (even to the physicists) mathematical formalism at the beginning of the century. There was no previous, heuristic understanding of the phenomena, and yet physics was *extracted* from the mathematical model. From then on, we have been trying to interrelate the physics and the mathematics of quantum theory into a full, convenient comprehension of the Universe.

§9.1 Præliminaris

In order to present a *taste* of the mathematical formalism of quantum mechanics, one has to introduce a few mathematical concepts that may seem unrelated to physics at first glance. As discussed before, quantum theory arose from this mathematical formalism, and it was unfamiliar even to the physicists at that time. It is the pinnacle of decades of research and theoretical refinement, and it has undergone exhaustive revision throughout the years. It should not be surprising that it seems so sombre at first.

As was said before, a physical process, e.g. a measurement, is often referred to as an *operator* in the language of quantum mechanics. The mathematical representation of such an operator is usually a matrix. Mathematically speaking, everything we can observe or measure can be represented by these extensive numerical arrays. Matrices can be so complex in structure so as to possess even an infinite amount of rows and columns; they belong in abstract spaces which, as one might just expect, are said to be *infinite-dimensional* spaces⁶³. Despite the common misuse of technical terms as *operator*, *matrix*, *space*, *dimensions*,

⁶¹Paul Davies, from the prologue to HEISENBERG, W. (1958) *Physics and Philosophy*. Penguin Books.

⁶²For a complete reference see: DE LA PEÑA, L. (2003). “Introducción a la Mecánica Cuántica.” Fondo de Cultura Económica.

⁶³The simplest matrix, a 1x1 matrix, is simply a number

etc, one has to keep in mind that no “metaphysical” connotations come along with them. They are simply mathematical objects with fancy names that are sometimes reminiscent of science fiction.

We say that the states a physical system can find itself in are elements that belong to very specific mathematical spaces called *Hilbert spaces*, in honour of the German mathematician David Hilbert⁶⁴. The geometry we need to use to describe Nature quantum-mechanically can be referred to as *quantum geometry*; it is the study of the shape and structure of these abstract spaces. How close or far away can things be in this kind of quantum structures is a concern of this branch of mathematics, usually called non-commutative geometry, as a reference to the way quantum-mechanical operators do not commute.

$$\begin{array}{ccc}
 \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} & \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} & \dots & \begin{pmatrix} 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & & \ddots \end{pmatrix} \\
 \text{2x2 matrix} & \text{3x3 matrix} & & \text{Infinite matrix}
 \end{array}$$

Figure 1.28: Examples of matrices

Matrices can have as many entries as we need. A matrix can possess all the information needed to describe a physical system. We begin by studying an oversimplification of the quantum-mechanical description of the nucleon proposed by Heisenberg, Condon, and Cassen in 1932⁶⁵. Here, we retake the study of the plane \mathbb{R}^2 from a quite different perspective, and describe it with the usual coordinates (x, y) . This time, however, we decompose this description by saying that $(x, y) = x(1, 0) + y(0, 1)$, where x and y are real numbers and the vectors can be represented both as rows or as columns, according to what is most comfortable in context. I.e. $\begin{pmatrix} x \\ y \end{pmatrix}$ and $(x \ y)$ mean exactly the same thing. We also define a *matrix-vector* multiplication by the following rules:

⁶⁴The notion of Hilbert Spaces generalises the notions of Euclidean Spaces; most importantly, it deals with the minutiae of possibly having infinite dimensions.

⁶⁵SRIVASTAVA, B. B., (2006), *Fundamentals of Nuclear Physics*, Rastogi Publications, India.

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax+by \\ cx+dy \end{pmatrix}$$

Figure 1.29: Rules of matrix-vector multiplication. The rows of the matrix “act” on the column vector and create a new vector. Here a, b, c, d , as well as x, y , are all real numbers.

$$\begin{pmatrix} a & b \\ \dots \end{pmatrix} \begin{pmatrix} x \\ y \\ \dots \end{pmatrix} = \begin{pmatrix} ax+by \\ \dots \end{pmatrix}$$

Figure 1.30: How the first row acts on the column; the result is the first entry of the new vector.

§9.2 Condon & Cassen: A Primitive Formalism

We now define protons and neutrons to be the fundamental elements of this two-dimensional space, i.e. $(1, 0)$ and $(0, 1)$, which from now on we only represent as column-vectors. The nucleon is then understood to be just a particular state of a particle that is simultaneously a neutron, and a proton when “no-one is watching,” but collapses into either one or the other in the precise moment when someone decides to *look* (measure). We carry on to identify these collapses with four mathematical operators that represent *verbatim* what we wish to portray.

$$\begin{aligned} |p\rangle &= \begin{pmatrix} 1 \\ 0 \end{pmatrix} & \hat{\pi}^- &= \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} & \hat{q}^- &= \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \\ |n\rangle &= \begin{pmatrix} 0 \\ 1 \end{pmatrix} & \hat{\pi}^+ &= \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} & \hat{q}^+ &= \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

Figure 1.31: Vector representation of the proton $|p\rangle$, the neutron $|n\rangle$, and the four quantum-mechanical operators: $\hat{\pi}^-$, $\hat{\pi}^+$, \hat{q}^- , and \hat{q}^+

As we can see by applying the rules of matrix-vector multiplication, each operator represents a physical process (Fig. 1.32). The first one shows how a proton is turned into a neutron; the second one how a neutron becomes a proton. The third operator annihilates the proton, and the last one annihilates the neutron. This kind of mathematical description is usually referred to as

matrix mechanics or *Jordan mechanics*, in honour of the German physicist Ernst P. Jordan (1902 - 1980), and is perfectly equivalent to the wave mechanics proposed by E. Schrödinger. With these (and much more complex) tools, it is possible to give a full description of how the Universe behaves at such scales, regardless of the human notions that may now seem more of an inconvenience than a helpful linguistic resource.

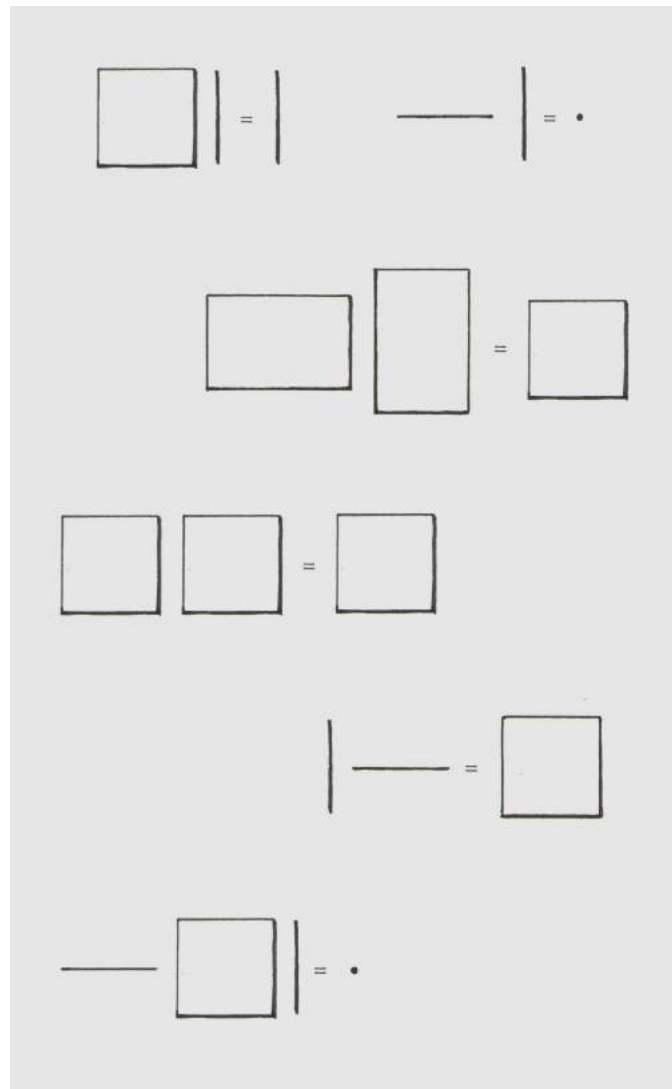
$$\begin{aligned} \hat{\Pi}^- |p\rangle &= \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} = |n\rangle & \hat{\Pi}^+ |n\rangle &= \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = |p\rangle \\ \hat{Q}^- |p\rangle &= \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} & \hat{Q}^+ |n\rangle &= \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{aligned}$$

Figure 1.32: Quantum-mechanical operators acting on the fundamental particles

At this point, all of the knowledge developed in the previous sections becomes essential to grasp fully some of the implications of quantum theory. A. Einstein, amongst many other physicists of his time, was opposed to the paradigm suggested by a strict interpretation of quantum mechanics. He believed the Universe *ought* to follow certain deterministic rules, and not be condemned to chance, probabilities, and randomness.

Chapter 2

An Overview of the Mathematics behind Quantum Theory



An Overview of the Mathematics behind Quantum Theory

Quantum mechanics is based on its mathematical formalism. Some physical theories, like electromagnetism or thermodynamics, were only formalised in the language of mathematics after a thorough understanding of the real, physical phenomena. Quantum theory, on the other hand, was built upon concepts that were previously anticipated by theoretical explorations, but not actually *seen* in physical situations. As opposed to any other formal physical model, the mathematics of quantum mechanics came first, and the actual physics came last.

Much of the beauty of quantum mechanics comes from its theoretical structure. As was mentioned in section §9 of chapter 1, quantum mechanics has two equivalent descriptions, namely *matrix mechanics*, and *wave mechanics*. There are essentially three major branches of mathematics that hold matrix mechanics together, linear algebra, group theory, and probability theory. Fourier analysis is the basis of wave mechanics, and the equivalence between these two versions allows us to understand quantum theory from both perspectives at once.

This section is dedicated to the sole purpose of providing an overview on selected topics that could allow a smooth transition into the formalities of quantum theory. Dirac's notation is extremely useful to do the actual calculations needed for a course on quantum physics, so it is better to introduce it and grow acquainted with it as soon as possible. The mathematical spaces and operators used in quantum physics can be quite abstract, but once its fundamentals have been understood at an elementary level, it is fairly simple to extend this knowledge to the level needed to handle all computations confidently. It is also important to acquire an intuition for the mathematical concepts dealt with in quantum theory.

§1 Vectors & Vector Spaces

Every quantum mechanical system has an associated wave function. Wave functions belong to a set of real or complex functions that has a *linear* structure. Linear structures are a wide and profoundly important topic in the fields of both mathematics and physics, the most fundamental of these are *vector spaces*. We begin by analysing the notions of vectors and vector spaces.

§1.1 What *is* a vector?

A *vector* is, roughly speaking, an element of a vector space. Our first encounter with vectors is usually related to the study of kinematics and dynamics, in Newtonian physics. Vectors are commonly defined as quantities with *magnitude* and *direction*, but there is much more to vectors than this. The most common way to introduce vectors is by describing the plane \mathbb{R}^2 , which is perhaps the simplest example of a vector space. Vector spaces are sets with an underlying structure. These sets are equipped with a binary operation called *addition*, defined for elements of the set; this means that two vectors can be combined into a new vector. Vectors can be *scaled* up or down, i.e. shortened or elongated, by means of a *scalar multiplication*¹.

But, intuitively speaking, what do we mean by *vectors* and *vector spaces*? The best way to think of a vector space is by recurring to the most basic, intuitive examples, i.e. \mathbb{R}^2 and \mathbb{R}^3 . Elements of \mathbb{R}^2 are usually thought of as “points” in the plane with coordinates $\begin{pmatrix} x \\ y \end{pmatrix}$, but we also find it useful to describe forces, velocities, accelerations, etc with arrows whose length represents a magnitude and whose coordinates point in the vector’s direction. For any practical purposes, these two notions are merely two ways of interpreting the exact same object, just as a cup can be used to drink coffee or tea, and one never thinks of it as two different objects.

So, whether we describe forces with vectors, whose magnitude and direction are encoded in the vector \vec{F} and its coordinates, or we think of an abstract object belonging to a structured set V called a vector space, we are talking about the exact same thing. However, we can distinguish both notions by means of notation, that way it is perfectly and unequivocally clear what is meant. I.e. we shall write

$$\vec{x} = \begin{pmatrix} x \\ y \end{pmatrix}$$

whenever we use vectors to describe arrows, i.e. quantities with length and direction, and

$$|x\rangle = \begin{pmatrix} x \\ y \end{pmatrix}$$

whenever we use vectors as abstract elements of a vector space. “ $|x\rangle$ ” is read *ket x*, whilst “ $\langle x|$ ” is read *bra x*; of course, “ $\langle x|y\rangle$ ” is the *bra(c)ket* of x and y .

¹Precise definitions for this and other algebraic structures are given in *Appendix 5: Reminders of Algebraic Definitions*

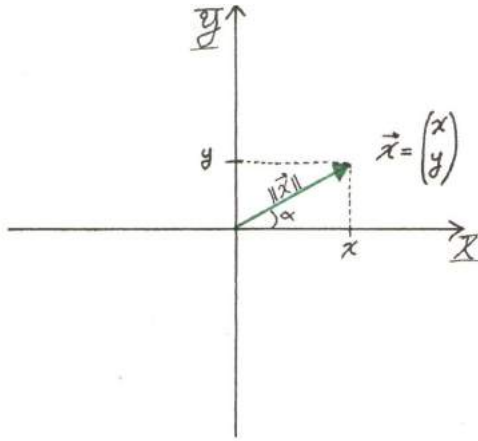


Figure 2.1: The plane \mathbb{R}^2 and a vector $\vec{x} \in \mathbb{R}^2$.

§1.2 The Pythagorean Theorem: Norm

A vector \vec{x} as the one depicted in Figure 2.1 has coordinates x and y ; its length, also called its *norm*, can be deduced from the Pythagorean theorem if we identify each coordinate as the side of a right-angled triangle with angle α . This means that the vector's norm is given by this triangle's hypotenuse, labelled $\|\vec{x}\|$ in the diagram. Therefore,

$$x^2 + y^2 = \|\vec{x}\|^2$$

where $\|\vec{x}\|$ is the vector's length (norm). If we define a *matrix multiplication* as we did in section §1.7, then the norm of \vec{x} (squared) can be expressed as

$$(x \ y) \begin{pmatrix} x \\ y \end{pmatrix} = x^2 + y^2$$

where the numerical array with the coordinates of \vec{x} is seen as a matrix. This is useful to define a more generalised expression for a vector's length. To see this, we can study the case of a vector $\vec{x} \in \mathbb{R}^3$

In this case, \vec{x} casts a shadow over the XY-plane that resembles the hypotenuse of a right-angled triangle with sides x and y , whose length is obviously $\sqrt{x^2 + y^2}$, also there is a standing right-angled triangle with sides z and the XY-shadow of \vec{x} , and hypotenuse $\|\vec{x}\|$. Therefore, the square of the norm of $\vec{x} \in \mathbb{R}^3$ is simply

$$(\sqrt{x^2 + y^2})^2 + z^2 = x^2 + y^2 + z^2 = \|\vec{x}\|^2$$

again by the Pythagorean theorem.

So, in general, we can say that the length (squared) of a vector $|x\rangle$ is given

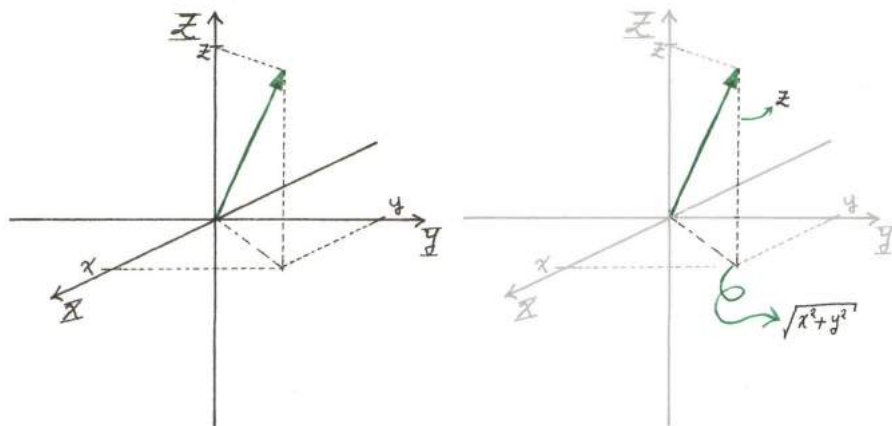


Figure 2.2: The norm of \vec{x} for a vector $\vec{x} \in \mathbb{R}^3$.

by the matrix multiplication

$$(x \ y \ z) \begin{pmatrix} x \\ y \\ z \end{pmatrix} = x^2 + y^2 + z^2$$

§1.3 Matrix Multiplication

The previous example shows the usefulness of multiplying matrices, and perhaps justifies the way we decided to define matrix multiplication. In general, any matrix is an $m \times k$ array (labelled \hat{A}) that can only be multiplied from the left to another $k \times n$ matrix \hat{B} , resulting in a $m \times n$ matrix, where $m, k, n \in \mathbb{N}$, i.e.

$$\hat{A}\hat{B} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mk} \end{pmatrix} \begin{pmatrix} b_{11} & \dots & b_{1n} \\ b_{21} & \dots & b_{2n} \\ \vdots & & \vdots \\ b_{k1} & \dots & b_{kn} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^k a_{1j}b_{j1} & \dots & \sum_{j=1}^k a_{1j}b_{jn} \\ \vdots & & \vdots \\ \sum_{j=1}^k a_{mj}b_{j1} & \dots & \sum_{j=1}^k a_{mj}b_{jn} \end{pmatrix}$$

where the multiplication rule is given for rows and columns as follows:

The first row of matrix \hat{A} is multiplied by the first column of matrix \hat{B} to obtain the first entry of matrix $\hat{A}\hat{B}$. This means that the first element of row one in \hat{A} is multiplied by the first element of column one in \hat{B} , and each of these pairs is added up to obtain the complete sum.

$$\hat{A}\hat{B}_{11} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \end{pmatrix} \begin{pmatrix} b_{11} \\ b_{21} \\ \vdots \\ b_{k1} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^k a_{1j}b_{j1} & \dots \\ \vdots & \end{pmatrix}$$

and the same for the rest of the entries

$$(\hat{A}\hat{B})_{il} = \begin{pmatrix} \vdots \\ a_{i1} & a_{i2} & a_{i3} & \dots & a_{ik} \\ \vdots \end{pmatrix} \begin{pmatrix} b_{1l} \\ b_{2l} \\ b_{3l} & \dots \\ \vdots \\ b_{kl} \end{pmatrix} = \begin{pmatrix} \vdots \\ \dots & \sum_{j=1}^k a_{ji}b_{jl} & \dots \\ \vdots \end{pmatrix}$$

This way, a vector $|x\rangle \in \mathbb{R}^3$ is also a 1×3 matrix that can be multiplied by itself to produce a scalar (a 1×1 matrix, if you will) corresponding to its length squared, as we did in the previous section.

It is important to notice that matrix multiplication is, in general, not commutative, i.e. $\hat{A}\hat{B} \neq \hat{B}\hat{A}$. Moreover, for some matrices $\hat{A}\hat{B}$ is defined, but $\hat{B}\hat{A}$ is not². Of course, \hat{A} always commutes with \hat{A} .

§1.4 Transposition

It might be useful then, to define a new matrix obtained by the “transposition” of vector $|x\rangle$, i.e.

$$|x\rangle = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \implies \langle x| = (x \quad y \quad z)$$

This way, the square of the length of vector $|x\rangle$ can be written in a much simpler form as

$$\langle x|x\rangle = (x \quad y \quad z) \begin{pmatrix} x \\ y \\ z \end{pmatrix} = x^2 + y^2 + z^2 = \|\vec{x}\|^2$$

And we can generalise the concept of transposition to any $m \times n$ matrix as follows³,

²What happens if \hat{A} is a $m \times k$ matrix, and \hat{B} is a $k \times n$ matrix, where $m \neq k \neq n$?

³One can notice that a different notation is used to transpose matrices than to transpose vectors, which are also matrices. This should not represent a problem; this is only a way to identify easily a vector from a matrix without explicitly writing its components. This means that $\langle x| = \vec{x}^t$

$$\hat{A}^t = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ \vdots & & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mk} \end{pmatrix}^t = \begin{pmatrix} a_{11} & \dots & a_{m1} \\ a_{12} & \dots & a_{m2} \\ \vdots & & \vdots \\ a_{1k} & \dots & a_{mk} \end{pmatrix}$$

One can imagine the matrix \hat{A} as being a sticker that can be removed from the page; the *transposition operation* would be to remove this “sticker” from the top-right corner, and replacing it backwards, leaving the bottom-left corner at the top-right.

§1.5 What to Imagine when Talking about \mathbb{R}^n

Most problems in physics involve more than three independent variables, and the corresponding graphs exceed the three dimensions one can visualise. In *Appendix 1: The Structure of Space and Time*, the notion of the *fourth dimension* is discussed in depth; it is not something one can draw or imagine, but fortunately we need not to. Geometrically speaking, it suffices with the analysis of *shadows*, projections from higher dimensions onto the space \mathbb{R}^3 ; algebraically, however, it is as simple as adding new coordinates to the spaces of interest. So, for instance, \mathbb{R}^4 corresponds to the vector space whose elements are

$$\begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} \in \mathbb{R}^4$$

and one can imagine a space similar to \mathbb{R}^3 , knowing there is an extra axis that points in a direction simultaneously perpendicular to all X , Y , and Z axes. Just as we draw \mathbb{R}^3 in a flat piece of paper by drawing two perpendicular axes and one tilted line representing the third direction, we can imagine the space \mathbb{R}^3 with a new tilted axis that represents the fourth direction, being always aware that it is, in fact, perpendicular to the rest of the axes.

In general, the space \mathbb{R}^n is just a vector space with n orthogonal⁴ axes and whose elements can be represented by n -dimensional arrays. We can also define the length of these vectors in \mathbb{R}^n by simply extending the Pythagorean theorem as we did from \mathbb{R}^2 to \mathbb{R}^3 . This should be justified by the fact that the “shadow” this vector casts over any two-dimensional plane will make a right-angled triangle with these two coordinates. So, for vectors in \mathbb{R}^4

$$\|\vec{x}\|^2 = \langle x|x \rangle = (x \ y \ z \ w) \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = x^2 + y^2 + z^2 + w^2$$

⁴orthogonal = perpendicular

And in general,

$$\|\vec{x}\|^2 = \langle x|x \rangle = (x_1 \quad x_2 \quad \dots \quad x_n) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \sum_{j=1}^n x_j^2$$

for any $|x\rangle \in \mathbb{R}^n$.⁵

§1.6 More on Matrix Multiplication

In general, two different vectors $|x\rangle, |y\rangle \in \mathbb{R}^n$ can be matrix-multiplied to obtain a scalar, using the transposition operation defined in §1.4, e.g.

$$\langle x|y \rangle = (x_1 \quad x_2 \quad \dots \quad x_n) \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \sum_{j=1}^n x_j y_j$$

This operation is called the *inner product* of $|x\rangle$ and $|y\rangle$. Notice also that matrix multiplication rules allow us to multiply these two vectors in the inverse order, and the resulting object is a $n \times n$ matrix

$$|x\rangle \langle y| = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} (y_1 \quad y_2 \quad \dots \quad y_n) = \begin{pmatrix} x_1 y_1 & x_1 y_2 & \dots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \dots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_n y_1 & x_n y_2 & \dots & x_n y_n \end{pmatrix}$$

The geometric meaning of these last two operations will be discussed later. As a mnemonic device, one can draw scalars, vectors, and the corresponding matrices that “act” on such vectors, as dots, lines and rectangles. Matrix multiplication rules are depicted on the cover of Part II.

§1.7 Inner Product and its Geometrical Meaning

Transposing a vector $|x\rangle$ and matrix-multiplying it to $|y\rangle$ from the left to obtain a scalar is a useful operation, not only to compute one vector’s length, but also to visualise, without the need of a graph, how any two vectors $|x\rangle$ and $|y\rangle$ interact geometrically. Any one vector lies on a one dimensional line; any two vectors lie, in general, on a two-dimensional plane, provided they are not a scaled version of the same vector. Three vectors span a three-dimensional space, and so on. To understand how two given vectors relate geometrically, we focus our attention to the case of $|x\rangle$ and $|y\rangle$ in a plane.

⁵This is why books on topology define the Euclidean norm as $\|\vec{x}\| = \sqrt{\sum_{j=1}^n x_j^2}$, which is simply an extension of the Pythagorean theorem.

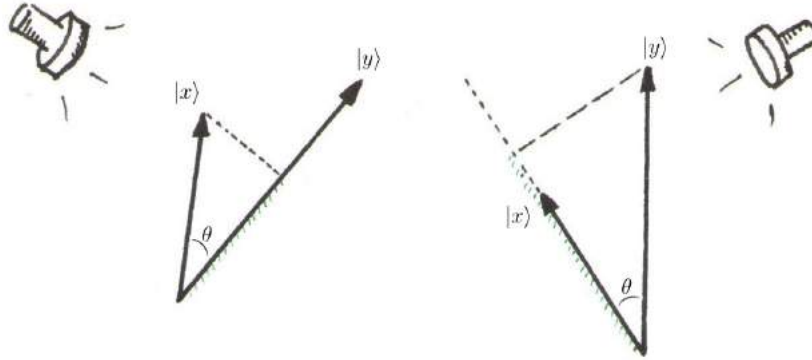


Figure 2.3: Inner product The vector $|x\rangle$ casts a shadow P_{xy} over $|y\rangle$. Analogously, $|y\rangle$ casts a shadow over the axis where $|x\rangle$ lies. If P_{xy} is the projection of $|x\rangle$ over $|y\rangle$, then P_{yx} is the projection of $|y\rangle$ over $|x\rangle$

Vectors $|x\rangle$ and $|y\rangle$ in Figure 2.3 are detached of any coordinate system; the origin could be very far away and rotated some unknown angle with respect to our perspective. Regardless, one could easily attach a new frame of reference in which, for example, vector $|y\rangle$ lies entirely on the X-axis. In that case, the new coordinates describing these vectors would be

$$|x\rangle = \begin{pmatrix} P_{xy} \\ x_2 \end{pmatrix} \quad \& \quad |y\rangle = \begin{pmatrix} \|\vec{y}\| \\ 0 \end{pmatrix}$$

where P_{xy} represents the *projection of $|x\rangle$ over $|y\rangle$* . Consequently, the inner product of $|x\rangle$ and $|y\rangle$ is

$$\langle x|y\rangle = (P_{xy} \quad x_2) \begin{pmatrix} \|\vec{y}\| \\ 0 \end{pmatrix} = P_{xy} \cdot \|\vec{y}\|$$

If a light-source is then placed “over” $|x\rangle$ as in Figure 2.3, it will cast a shadow P_{xy} perpendicularly onto $|y\rangle$ so that it forms a right-angled triangle. From the definition of $\cos(\theta)$ one can readily see that the “shadow” of $|x\rangle$ over $|y\rangle$ corresponds precisely to the adjacent side of the angle θ , and the norm of $|x\rangle$ corresponds to the triangle’s hypotenuse, i.e.

$$\cos(\theta) = \frac{\text{adjacent}}{\text{hypotenuse}} = \frac{P_{xy}}{\|\vec{x}\|}$$

From these last two equations it is fairly uncomplicated to prove that

$$P_{xy} = \|\vec{x}\| \cdot \cos(\theta) \quad \& \quad P_{xy} = \frac{\langle x|y\rangle}{\|\vec{y}\|}$$

which implies that $\langle x|y\rangle = \|\vec{x}\|\|\vec{y}\| \cdot \cos(\theta)$, where $\langle x|y\rangle = \sum_{j=1}^n x_j y_j$ (as was defined above), θ is the angle between the two vectors, $\|\vec{x}\| = \sqrt{\langle x|x\rangle}$, and $\|\vec{y}\| = \sqrt{\langle y|y\rangle}$.

To handle confidently the following relations might be useful hereafter, and they should be understood as follows:

$$P_{xy} = \frac{\langle x|y\rangle}{\|\vec{y}\|}$$

is the projection of $|x\rangle$ on $|y\rangle$

$$P_{yx} = \frac{\langle x|y\rangle}{\|\vec{x}\|}$$

is the projection of $|y\rangle$ on $|x\rangle$

$$\langle x|x\rangle = \|\vec{x}\|^2$$

is the norm of $|x\rangle$ squared

The value of these last three equations can hardly be overstated, first because they emphasise the importance of the *transposition* operation, by means of which a vector $|x\rangle$ is transformed into $\langle x|$ to understand the way it interacts with another vector $|y\rangle$. These equations have no specific reference to a particular vector space V , so their validity can be extended to *any* vector space that allows the definition of an inner product, however abstract it may be.

§2 Linear Transformations

One particular case of matrix multiplication that is especially important for both physics and mathematics is that of a vector multiplied from the left by a matrix, e.g.

$$\hat{A}|x\rangle = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax + by \\ cx + dy \end{pmatrix}$$

as was shown in section §9 of chapter 1. The result is another vector $|y\rangle \in V$, in this case \mathbb{R}^2 . Matrix \hat{A} must forcefully have as many columns as vector $|x\rangle$ has entries, otherwise the resulting vector would not be properly defined. The geometric meaning of this new vector $|y\rangle = \hat{A}|x\rangle$ can be seen in Figure 2.4. The new vector $|y\rangle$ is a scaled and/or rotated version of the original vector $|x\rangle$. This means that \hat{A} *transformed* $|x\rangle$ by means of a scalar multiplication and a rotation.

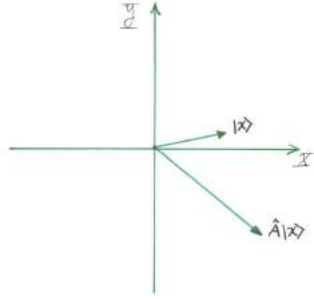


Figure 2.4: Geometric meaning of $|x\rangle$, and the way it is transformed by \hat{A} into a new vector $\hat{A}|x\rangle$

A *linear transformation* is a function, an operation that preserves the so-called linear structure of the vector space. The matrix \hat{A} in the previous example represents precisely what a linear transformation is. We defined vector spaces as sets whose elements can be added and scaled, so an operation that preserves this structure should allow the resulting elements to be also added and scaled. We begin by studying the algebraic and geometric structure of vector addition.

§2.1 Basis Vectors

Any vector $|x\rangle$ is uniquely defined by its coordinates, provided one establishes a fixed frame of reference. In the case of \mathbb{R}^2 , it is the pair of components x and y that define the vector unequivocally. These two scalars can be interpreted as an elongation of two basis vectors $|e_1\rangle$ and $|e_2\rangle$. I.e.

$$|x\rangle = \begin{pmatrix} x \\ y \end{pmatrix} = x \begin{pmatrix} 1 \\ 0 \end{pmatrix} + y \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Or simply $|\mathbf{x}\rangle = x|e_1\rangle + y|e_2\rangle$, provided we define the following: ⁶

$$|e_1\rangle := \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \& \quad |e_2\rangle := \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Notice that $|x\rangle$ is a vector, an element of the vector space V , not to be confused with the scalar x , a real number that *scales* the basis vector $|e_1\rangle$, and provides the first component of the matrix representation $\begin{pmatrix} x \\ y \end{pmatrix}$.

Of course, a vector $|x\rangle \in \mathbb{R}^3$ can also be decomposed into the basis vectors of \mathbb{R}^3 , i.e. $|e_1\rangle$, $|e_2\rangle$, and $|e_3\rangle$, i.e.

$$|x\rangle = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = x \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + y \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + z \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Or $|\mathbf{x}\rangle = x|e_1\rangle + y|e_2\rangle + z|e_3\rangle$, where

$$|e_1\rangle := \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad |e_2\rangle := \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \& \quad |e_3\rangle := \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

⁶The symbol $:=$ means the quantity on the left is given as a definition

Notice how $|e_1\rangle$ as a basis element of \mathbb{R}^2 is totally different from $|e_1\rangle$ as a basis element of \mathbb{R}^3 . Every vector space V has its own set of basis elements, and every vector $|v\rangle \in V$ can be expressed unambiguously as a combination of these basis elements. This combination is called a *linear combination*. The number of basis vectors needed to describe vectors $|v\rangle \in V$ is called the *dimension* of V .

Of course, vectors $|x\rangle \in \mathbb{R}^4$ are decomposed as linear combinations of basis elements as follows,

$$|x\rangle = \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = x \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + y \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + z \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} + w \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

Or $|\mathbf{x}\rangle = x|e_1\rangle + y|e_2\rangle + z|e_3\rangle + w|e_4\rangle$, where $\{|e_j\rangle\}_{j=1}^4$ is the set of basis vectors.

So, \mathbb{R}^2 is a vector space of dimension 2, and \mathbb{R}^3 is a vector space of dimension 3; evidently, \mathbb{R}^n is a vector space of dimension n , since every element $|x\rangle \in \mathbb{R}^n$ is such that

$$|x\rangle = \sum_{j=1}^n \lambda_j |e_j\rangle$$

where each λ_j represents the coordinate in the j^{th} direction. Despite the fact that one cannot imagine these vectors, it suffices with their algebraic expansion as a linear combination of the set of basis vectors $\{|e_j\rangle\}_{j=1}^n$ to operate and calculate anything related to such abstract spaces.

§2.2 Linearity

If $|v\rangle$ and $|u\rangle$ are vectors in a vector space V , then the scaling $\lambda|v\rangle$ or $\xi|u\rangle$ are also vectors in V , provided $\lambda, \xi \in \mathbb{R}$; of course, $|w\rangle = \lambda|v\rangle + \xi|u\rangle$ is also a vector in V .

Then, if \hat{T} is a linear transformation, \hat{T} applied to the sum of $|v\rangle$ and $|u\rangle$ should be the sum of the individual transformations, provided one wishes to preserve the linear structure of the space V , i.e. the ability to scale and add elements of the space. Scaling $|v\rangle$ by the factor λ and then transforming it by \hat{T} should be equivalent to applying \hat{T} first, and then scaling it by λ . This is precisely the definition of linearity for a given transformation \hat{T} .

To be precise, a linear transformation from a space into itself⁷ is a $\hat{T} : V \rightarrow V$ such that $|v\rangle \in V$ is transformed into another vector⁸ $\hat{T}|v\rangle$, also in V , which means that

$$\hat{T}(|v\rangle + |u\rangle) = \hat{T}|v\rangle + \hat{T}|u\rangle$$

⁷A linear function can transform vectors from one space into another, not necessarily the same one. We now focus on the former particular case.

⁸In traditional notation, the vector \vec{x} would be transformed by the linear transformation f as $f(\vec{x})$, instead of something like $\hat{F}|x\rangle$.

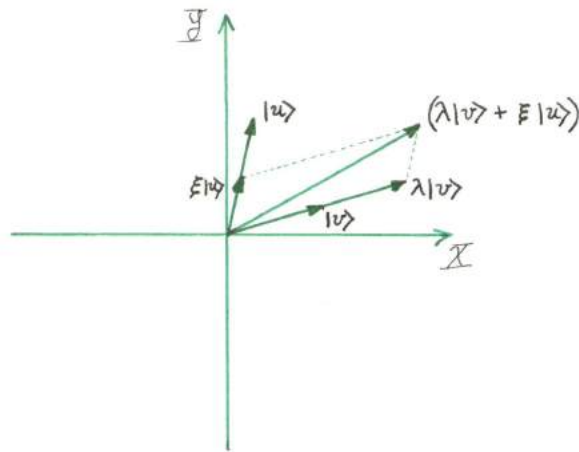


Figure 2.5: The geometric meaning of vector addition. Vector $|v\rangle$ is scaled by the factor λ , whilst vector $|u\rangle$ is scaled by ξ . Adding these two vectors is equivalent to a displacement from the origin to the tip of $\lambda|v\rangle$, and then $\xi|u\rangle$ from the resulting point. The new vector $(\lambda|v\rangle + \xi|u\rangle)$ is the diagonal of the parallelogram with sides $\lambda|v\rangle$ and $\xi|u\rangle$.

Also,⁹

$$\hat{T}(\lambda|v\rangle) = \lambda(\hat{T}|v\rangle)$$

⁹This can be proved for vectors in \mathbb{R}^2 as follows: $\hat{A}(|x\rangle + |y\rangle) =$
 $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \end{pmatrix} = \begin{pmatrix} a(x_1 + y_1) + b(x_2 + y_2) \\ c(x_1 + y_1) + d(x_2 + y_2) \end{pmatrix}$
 $= \begin{pmatrix} ax_1 + bx_2 + ay_1 + by_2 \\ cx_1 + dx_2 + cy_1 + dy_2 \end{pmatrix} = \begin{pmatrix} (ax_1 + bx_2) + (ay_1 + by_2) \\ (cx_1 + dx_2) + (cy_1 + dy_2) \end{pmatrix}$
 $= \begin{pmatrix} (ax_1 + bx_2) \\ (cx_1 + dx_2) \end{pmatrix} + \begin{pmatrix} (ay_1 + by_2) \\ (cy_1 + dy_2) \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} =$
 $\hat{A}|x\rangle + \hat{A}|y\rangle$

The scaling property is quite easy to prove as well.

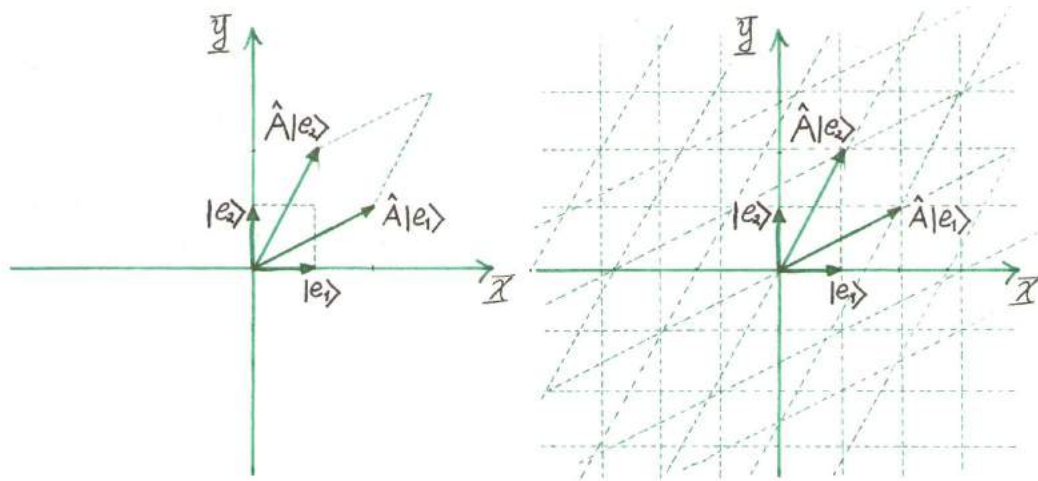


Figure 2.6: (Left) The linear transformation \hat{A} acting on the basis vectors $|e_1\rangle$ and $|e_2\rangle$. Any vector $|x\rangle$ is transformed by \hat{A} in a linear form; this implies that $\hat{A}|x\rangle = \hat{A}(x|e_1\rangle + y|e_2\rangle) = x(\hat{A}|e_1\rangle) + y(\hat{A}|e_2\rangle)$. Thus, knowing how \hat{A} affects the basis vectors implies knowing how \hat{A} affects any vector.

(Right) The vector space \mathbb{R}^2 can be fully described by the grid spanned by $|e_1\rangle$ and $|e_2\rangle$; equivalently, \mathbb{R}^2 can be described by the grid spanned by $\hat{A}|e_1\rangle$ and $\hat{A}|e_2\rangle$. This means that $\hat{A}|e_1\rangle$ and $\hat{A}|e_2\rangle$ are also a basis for \mathbb{R}^2 .

§2.3 To Illustrate...

Linear functions can be classified by means of the way they transform the basis vectors. A few examples might be useful to illustrate this point. A linear transformation \hat{A} as the one depicted in Figure 2.6 distorts the unit square spanned by the two basis vectors $|e_1\rangle$ and $|e_2\rangle$ (each of length 1) into a new parallelogram with sides $\hat{A}|e_1\rangle$ and $\hat{A}|e_2\rangle$, with a new area corresponding to the span of these two new vectors.

In this particular case (Figure 2.6), $\hat{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, so that $\hat{A}|e_1\rangle = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ and $\hat{A}|e_2\rangle = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$.

For any given linear transformation \hat{A} that distorts the usual grid of \mathbb{R}^2 into a new grid, the area of the new parallelogram is a good criterion to classify the amount of “distortion” this transformation induces. Any linear transformation \hat{L} that preserves the value of the area spanned by any two vectors, despite distortions, is classified into a group under the name of *special linear* transformations, as long as their relative orientation with respect to each other is also preserved. For the space \mathbb{R}^2 of dimension 2, the group is abbreviated as $SL(2)$.

Rotations

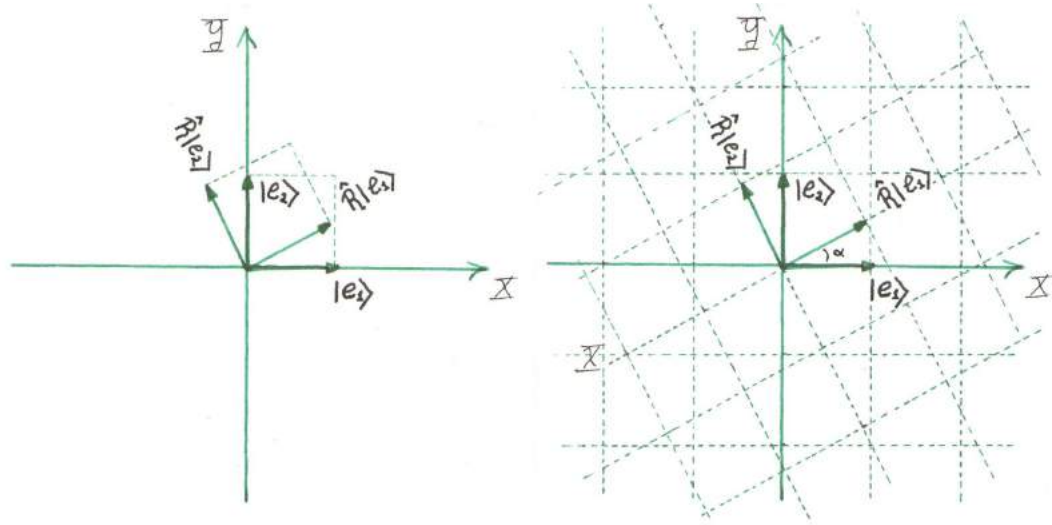


Figure 2.7: (Left) The linear transformation \hat{R} represents a rotation by an angle α . This means that for any vector $|v\rangle \in \mathbb{R}^2$, \hat{R} rotates $|v\rangle$ into a new vector $\hat{R}|v\rangle$. \hat{R} is a linear transformation, so it suffices to know how \hat{R} transforms the basis vectors in order to deduce how it transforms any other vector. (Right) The new grid spanned by $\hat{R}|e_1\rangle$ and $\hat{R}|e_2\rangle$.

The rotation \hat{R} by an angle α does not distort the area spanned by $|e_1\rangle$ and $|e_2\rangle$, thus it is said rotations are *rigid* transformations. It is linear since the rotation of the sum of two scaled vectors is the same as the sum of the two vectors individually rotated and scaled, i.e.

$$\hat{R}(\lambda|x\rangle + \xi|y\rangle) = \lambda(\hat{R}|x\rangle) + \xi(\hat{R}|y\rangle)$$

Linear transformations that preserve the area spanned by any two vectors, their relative orientation with respect to each other, and the angle between them, are classified into a group under the name of *special orthogonal* transformations. For the space \mathbb{R}^2 of dimension 2, the group is abbreviated as $SO(2)$.

Reflections

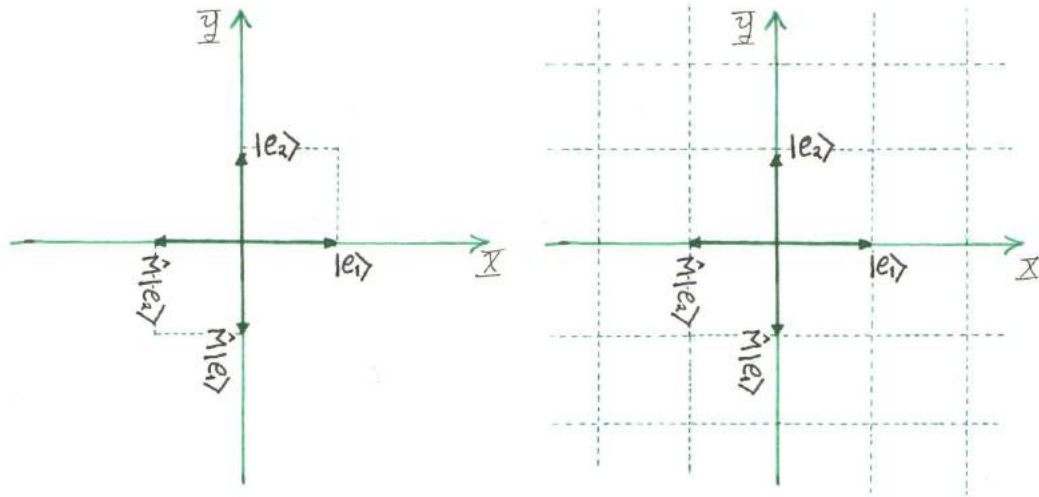


Figure 2.8: (Left) The linear transformation \hat{M} represents a reflection through a line passing by the origin at an angle of -45° . This means that for any vector $|v\rangle \in \mathbb{R}^2$, \hat{M} transforms $|v\rangle$ into a new mirror version of itself $\hat{M}|v\rangle$. (Right) The new grid spanned by $\hat{M}|e_1\rangle$ and $\hat{M}|e_2\rangle$ matches the original grid.

The reflection \hat{M} through this line by the origin does not distort the area spanned by $|e_1\rangle$ and $|e_2\rangle$, it does, however, reverse the orientation of any two vectors in \mathbb{R}^2 . Notice how $\hat{A}|e_1\rangle$ is always to the right of $\hat{A}|e_2\rangle$ in all other previous examples; in this case, however, the resulting vectors of $\hat{M}|e_1\rangle$ and $\hat{M}|e_2\rangle$ have an inverted orientation. It is linear since the reflection of the sum of two scaled vectors is the same as the sum of the two vectors individually reflected and scaled.

Linear transformations that preserve the area spanned by any two vectors and the angle between them, but fail to preserve their relative orientation with respect to each other, are classified into a group under the name of *orthogonal* transformations. For the space \mathbb{R}^2 of dimension 2, the group is abbreviated as $O(2)$.

§2.4 Examples in \mathbb{R}^3

Rotations

A rotation in \mathbb{R}^3 transforms the basis vectors $|e_1\rangle$, $|e_2\rangle$, and $|e_3\rangle$ into new vectors $\hat{R}|e_1\rangle$, $\hat{R}|e_2\rangle$, and $\hat{R}|e_3\rangle$. In this case, the volume spanned by the three basis vectors is a cube with sides of length 1, and thus $V = 1$. When transformed by this rotation \hat{R} , this volume is preserved, and the resulting grid spans the whole three-dimensional space, analogous to what happened with the rotation of \mathbb{R}^2 in Figure 2.7. Figure 2.9 depicts how the cube spanned by the three basis vectors is transformed under the action of \hat{R} , a three-dimensional rotation.

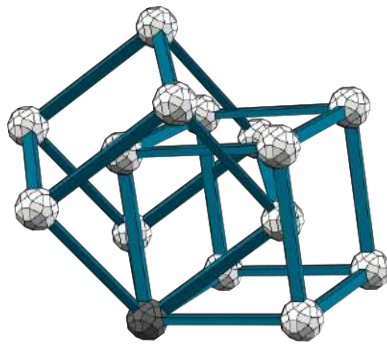


Figure 2.9: A transformation $\hat{R} \in SO(3)$

Linear transformations that preserve the volume spanned by any three vectors in \mathbb{R}^3 , their relative orientation with respect to each other, and the angle between them, are classified into a group under the name of *special orthogonal* transformations. For the space \mathbb{R}^3 of dimension 3, the group is abbreviated as $SO(3)$, to distinguish it from $SO(2)$.

General Transformations

In the case of a linear transformation \hat{D} that distorts the basis vectors $|e_1\rangle$, $|e_2\rangle$, and $|e_3\rangle$ into a flat surface, thus “collapsing” at least one of the three dimensions, \hat{D} is said to be *degenerate*. There are different degrees of degeneracy, according to how many dimensions are “collapsed” during the transformation, but whenever a transformation \hat{A} is non-degenerate, we classify it into the group of *general linear* transformations, abbreviated $GL(2)$, for transformations of the vector space \mathbb{R}^2 , and $GL(3)$ in the case of \mathbb{R}^3 .

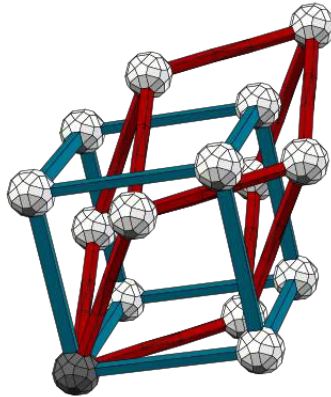


Figure 2.10: A linear transformation $\hat{L} \in GL(3)$ that transforms the cube spanned by the set of basis vectors into a new *parallelepiped*. The set $GL(3)$ of non-degenerate functions consists of all linear transformations that take the basis vectors $|e_1\rangle$, $|e_2\rangle$, and $|e_3\rangle$ into a new set of vectors $\hat{L}|e_1\rangle$, $\hat{L}|e_2\rangle$, and $\hat{L}|e_3\rangle$ that span a non-zero volume.

Special Transformations

When the linear transformation \hat{L} distorts the cube of basis vectors in such a way that the vectors per se and the angles between them are not preserved, but the volume spanned by the old and new vectors is, we classify \hat{L} into the group of *special linear* transformations, abbreviated $SL(3)$ for the case of \mathbb{R}^3 .

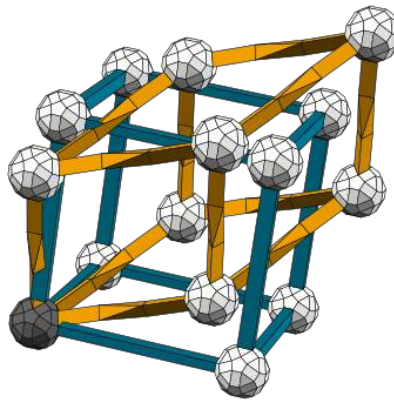


Figure 2.11: A linear transformation $\hat{L} \in SL(3)$ that preserves volume, despite re-orienting the three basis vectors $|e_1\rangle$, $|e_2\rangle$, and $|e_3\rangle$.

§2.5 The Determinant: A Useful Criterion I

The brief discussions in the previous examples exhibit the importance of *area* and *volume* to classify how linear transformations change the vector spaces they act upon. As long as the linear transformation $\hat{A} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is non-degenerate, the resulting vectors $\hat{A}|e_1\rangle$ and $\hat{A}|e_2\rangle$ span a parallelogram as the one in Figure 2.12. In the case of $\hat{A} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$, every non-degenerate transformation \hat{A} transforms the three basis vectors into a new set $\hat{A}|e_1\rangle$, $\hat{A}|e_2\rangle$, and $\hat{A}|e_3\rangle$ that span a parallelepiped as the ones depicted in Figures 2.9, 2.10, and 2.11.

In general, a parallelogram covers an area $A = \text{base} \cdot \text{height}$. To find the area spanned by the two vectors $\hat{A}|e_1\rangle$ and $\hat{A}|e_2\rangle$ we will need two auxiliary elements. First we draw a new vector $(\hat{A}|e_1\rangle)_{\text{rotated}90^\circ}$ by rotating $\hat{A}|e_1\rangle$ 90° to the left. Then we identify the new angle θ_2 , which is complementary to θ_1 ,

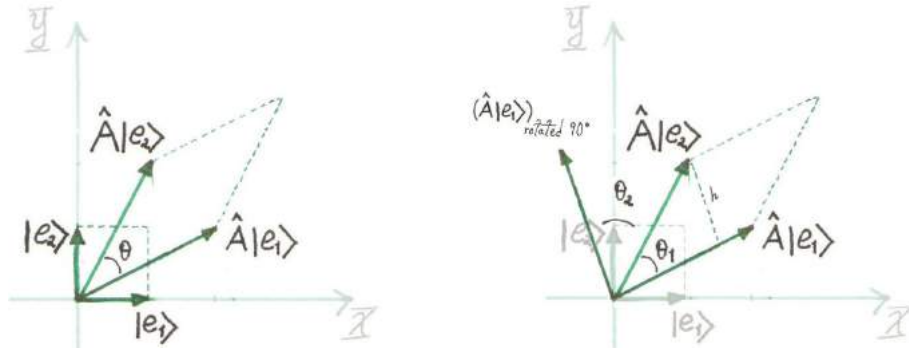


Figure 2.12: (Left) Parallelogram formed by $\hat{A}|e_1\rangle$ and $\hat{A}|e_2\rangle$. (Right) Auxiliary vector $(\hat{A}|e_1\rangle)_{rotated90^\circ}$ and angles θ_1 and θ_2 .

i.e. $\theta_1 + \theta_2 = 90^\circ$. The base of this parallelogram corresponds to the norm of $\hat{A}|e_1\rangle$, and the height to the vertical projection of $\hat{A}|e_2\rangle$ onto $\hat{A}|e_1\rangle$ (as defined in §1.7), which can be deduced from the definition of $\sin(\theta)$, i.e.

$$\sin(\theta_1) = \frac{\text{opposite}}{\text{hypotenuse}} = \frac{\text{height}}{\|\hat{A}|e_2\rangle\|}$$

which implies that

$$\text{height} = \|\hat{A}|e_2\rangle\| \cdot \sin(\theta_1)$$

and thus,

$$A = \text{base} \cdot \text{height} = \|\hat{A}|e_1\rangle\| \cdot \|\hat{A}|e_2\rangle\| \cdot \sin(\theta_1)$$

However, $\|\hat{A}|e_2\rangle\| \cdot \sin(\theta_1)$ can be expressed in a different and much more recognisable form if we identify the following facts:

First, for any pair of complementary angles, $\sin(\alpha) = \cos(90^\circ - \alpha)$, which can be seen by drawing a right-angled triangle and using the definition of $\sin(\alpha)$ and $\cos(\beta)$ for the two non-right angles. This means that $\sin(\theta_1) = \cos(\theta_2)$.

The size of $\hat{A}|e_1\rangle$ is not changed by the 90° rotation, i.e.

$$\|\hat{A}|e_1\rangle\| = \|(\hat{A}|e_1\rangle)_{rotated90^\circ}\|$$

So from these two facts we can conclude that

$$A = \text{base} \cdot \text{height} = \|\hat{A}|e_1\rangle\| \cdot \|\hat{A}|e_2\rangle\| \cdot \sin(\theta_1) = \|(\hat{A}|e_1\rangle)_{rotated90^\circ}\| \cdot \|\hat{A}|e_2\rangle\| \cdot \cos(\theta_2)$$

which is precisely the inner product of $(\hat{A}|e_1\rangle)_{rotated90^\circ}$ and $\hat{A}|e_2\rangle$. So, if we compute

$$\hat{A}|e_1\rangle = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ c \end{pmatrix}$$

we can identify that $(\hat{A}|e_1\rangle)_{rotated90^\circ}$ has coordinates $(\hat{A}|e_1\rangle)_{rotated90^\circ} = \begin{pmatrix} -b \\ a \end{pmatrix}$, which can be easily seen by rotating the whole plane 90° to the right and now drawing the vector's x -coordinate in the Y -axis, and its y -coordinate in the X -axis with the proper sign adjustment. In the case of $\hat{A}|e_2\rangle$, its coordinates are given by

$$\hat{A}|e_2\rangle = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} c \\ d \end{pmatrix}$$

Which means that the area spanned by both vectors is exactly

$$A = \|(\hat{A}|e_1\rangle)_{rotated90^\circ}\| \cdot \|\hat{A}|e_2\rangle\| \cdot \cos(\theta_2) = (-b \ a) \begin{pmatrix} c \\ d \end{pmatrix} = -bc + ad$$

This suggests we can define a new quantity $(ad - bc)$ for the matrix \hat{A} , since the area spanned by the two new vectors only depends on the four parameters a, b, c , and d . We define the *determinant* of matrix \hat{A} as

$$\det(\hat{A}) = \det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$$

and we identify $\det(\hat{A})$ as the area spanned by the two vectors $\begin{pmatrix} a \\ b \end{pmatrix}$ and $\begin{pmatrix} c \\ d \end{pmatrix}$.

This fact can be extended to any arbitrary vector space of dimension n , so we can reformulate our criterion to classify linear transformations according to the area or volume spanned by the transformation of the basis vectors with the aid of our new mathematical tool, the determinant.

§2.6 More on Vector Transposition

We recall the importance of vector transposition to compute norms. A vector $|x\rangle$ was transposed into $\langle x|$ to be matrix multiplied from the left to produce a scalar, namely the vector's norm squared. So, what happens when we compute the norm of a vector that has been transformed by the matrix \hat{A} . In the case of \mathbb{R}^2 , we find that

$$\left(\hat{A}|x\rangle\right)^t = \left[\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}\right]^t = \left[\begin{pmatrix} ax + by \\ cx + dy \end{pmatrix}\right]^t = (ax + by \quad cx + dy)$$

but this vector corresponds precisely to

$$(ax + by \quad cx + dy) = (x \ y) \begin{pmatrix} a & c \\ b & d \end{pmatrix} = \langle x| \hat{A}^t$$

This means that,

$$\left(\hat{A}|x\rangle\right)^t = \langle x| \hat{A}^t$$

and thus

$$\|\hat{A}|x\rangle\| = \langle x|\hat{A}^t\rangle\hat{A}|x\rangle = \langle x|\hat{A}^t\hat{A}|x\rangle$$

corresponds to the norm of $\hat{A}|x\rangle$.¹⁰ One can generalise this operation for any two vectors and a matrix \hat{A} as follows:

$$\langle x|\hat{A}|y\rangle = (x_1 \quad y_1) \begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}$$

which is simply a real number.

§2.7 The Determinant: A Useful Criterion II

We now summarise and reformulate the discussions given in sections §2.3 and §2.4 of this chapter in terms of determinants.

$M_{n \times n}(\mathbb{R})$

The set of all possible $n \times n$ matrices \hat{M} with real entries corresponds to matrices of the form

$$\hat{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

where a , b , c , and d are all real numbers¹¹. This includes matrices that span the set of basis vectors into a parallelogram of area $A = 0$ for the case of \mathbb{R}^2 , and a parallelepiped of volume $V = 0$ for the case of \mathbb{R}^3 , i.e. the set of degenerate matrices.

We concluded in section §2.5 that the geometrical meaning of $\det(\hat{A})$ corresponds to the area (or volume) that results from the transformation \hat{A} . That being the case, $M_{n \times n}(\mathbb{R})$ contains matrices with arbitrary determinants, i.e. $\det(\hat{A})$ can be any real number, including *zero*, corresponding to degenerate transformations.

¹⁰Explicitly,

$$\|\hat{A}|x\rangle\| = \langle x|\hat{A}^t\hat{A}|x\rangle = (x \quad y) \begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

¹¹The case of $M_{n \times n}(\mathbb{C})$ will be discussed in a further section.

GL(2) and GL(3)

If we decide to filter for non-degeneracy, we will define a new set of linear transformations \hat{L} such that the volume spanned by the resulting vectors of \hat{L} is non-zero. As was stated before, this set is referred to as the *general linear* group, $GL(2)$ for the case of \mathbb{R}^2 , and $GL(3)$ for the case of \mathbb{R}^3 . I.e.

$$\hat{L} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

such that $\det(\hat{L}) \neq 0$.

SL(2) and SL(3)

Linear transformations \hat{S} that preserve the volume and orientation spanned by the set of basis vectors $\{|e_j\rangle\}$, for $j = 2$ or $j = 3$, are part of a group of matrices called *special linear* transformations. As a consequence, these linear transformations preserve the area or volume spanned by *any* set of vectors. Since, in particular, the area or volume spanned by the basis vectors is preserved, the determinant of this kind of transformations should always be exactly 1. As was stated before, this set is abbreviated $SL(2)$ for the case of \mathbb{R}^2 , and $SL(3)$ for the case of \mathbb{R}^3 . I.e.

$$\hat{S} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

such that $\det(\hat{S}) = 1$.

SO(2) and SO(3)

The set of linear functions \hat{U} that preserve the area or volume spanned by the basis vectors, but also preserve the angle between them, and their relative orientation is classified as the group of *special orthogonal* transformations, abbreviated $SO(2)$ for the case of \mathbb{R}^2 , and $SO(3)$ for the case of \mathbb{R}^3 . The condition of preserving area or volume is already given by $\det(\hat{U}) = 1$, but the second condition cannot be expressed in terms of determinants.

However, we know that there is information regarding the angle θ between any two vectors in their inner product. Specifically,

$$\langle x|y \rangle = \|\vec{x}\| \cdot \|\vec{y}\| \cdot \cos(\theta) \quad \implies \quad \cos(\theta) = \frac{\langle x|y \rangle}{\|\vec{x}\| \cdot \|\vec{y}\|}$$

Since $\det(\hat{U}) = 1$ implies that norms are preserved, we need $\cos(\theta)$, and thus $\langle x|y \rangle$, to be preserved as well. This means that for every pair of vectors $|x\rangle$ and $|y\rangle$ their inner product $\langle x|y \rangle$ is the same as that of $\hat{U}|x\rangle$ and $\hat{U}|y\rangle$, i.e.

$$\langle x|y \rangle = (\langle x|\hat{U}^t) \hat{U}|y \rangle = \langle x|\hat{U}\hat{U}^t|y \rangle$$

$$\implies \hat{U}\hat{U}^t = \mathbb{1}$$

where $\mathbf{1}$ is the identity matrix ¹².

So the special feature for any linear transformation $\hat{U} \in SO(n)$ is that $\hat{U}\hat{U}^t = \mathbf{1}$ and $\det(\hat{U}) = 1$.

$O(2)$ and $O(3)$

A special instance occurs when a linear transformation \hat{O} is such that $\det(\hat{O}) = -1$. In this case, the orientation of the vector space V is changed by the linear transformation, as in the reflection example of Figure 2.8, §2.3.

The set of all linear transformations that preserve areas, or volumes in the case of \mathbb{R}^3 , and angles between vectors, but reverse the orientation of V , belong to the group of *orthogonal* transformations, labelled $O(2)$ and $O(3)$ respectively. Obviously, $SO(2) \subset O(2)$ and $SO(3) \subset O(3)$, so the defining conditions for $\hat{O} \in O(n)$ are $\hat{O}\hat{O}^t = \mathbf{1}$ and $\det(\hat{O}) = \pm 1$.

§2.8 Coordinate Transformations

Notice how a vector $|x\rangle \in \mathbb{R}^2$ can be expressed as a linear combination of the usual basis elements $|e_1\rangle$ and $|e_2\rangle$, but also as a combination of the new basis $\hat{A}|e_1\rangle$ and $\hat{A}|e_2\rangle$, as can be seen in Figure 2.13. The coordinates that describe $|x\rangle$ uniquely and unambiguously are determined by the frame of reference once chooses to work with. Whilst coordinates are artificial to a certain extent, the vector per se is unaltered by the way one decides to describe it.

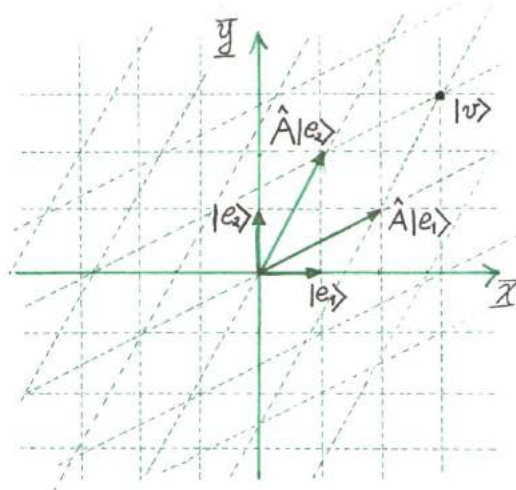


Figure 2.13: Coordinate transformation.

¹²The matrix equivalent of the real number 1,
 $\mathbf{1}|x\rangle = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} = |x\rangle$

The point $|v\rangle \in V$ can be described from the frame of reference of $|e_1\rangle$ and $|e_2\rangle$, where it has the pair of coordinates $|v\rangle = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$, or from that of $\hat{A}|e_1\rangle$ and $\hat{A}|e_2\rangle$, where its coordinates are given by $|v\rangle = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

There are two important facts about the way a linear function affects coordinates that one has to handle confidently in general. We concluded in §2.2, Figure 2.6 that knowing how a linear transformation \hat{A} affects the basis vectors implies knowing how \hat{A} affects any vector. $\hat{A}|x\rangle = \hat{A}(x|e_1\rangle + y|e_2\rangle) = x(\hat{A}|e_1\rangle) + y(\hat{A}|e_2\rangle)$. Thus, for any linear transformation,

$$\hat{A}|e_1\rangle = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ c \end{pmatrix} \quad \& \quad \hat{A}|e_2\rangle = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} b \\ d \end{pmatrix}.$$

In the case of \mathbb{R}^3 ,

$$\hat{A}|e_1\rangle = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ d \\ g \end{pmatrix} \quad \hat{A}|e_2\rangle = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} b \\ e \\ h \end{pmatrix}$$

&

$$\hat{A}|e_3\rangle = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} c \\ f \\ i \end{pmatrix}.$$

From this we can infer that the columns of any matrix \hat{A} consist precisely of the resulting coordinates of the transformation \hat{A} applied to the basis vectors¹³.

Suppose we have a vector $|v\rangle \in \mathbb{R}^2$ as the one depicted in Figure 2.13. As was discussed above, someone trying to describe $|v\rangle$ in terms of the standard, canonical basis of \mathbb{R}^2 will say it has coordinates $\begin{pmatrix} 3 \\ 3 \end{pmatrix}$, as opposed to someone describing it from the perspective of the new basis $\hat{A}|e_1\rangle$ and $\hat{A}|e_2\rangle$, where its coordinates are simply $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

This suggests that we need a method to translate any pair of coordinates from one frame of reference to the other, a pair of transformations \hat{T} and its inverse \hat{T}^{-1} that change $|v\rangle$, described in terms of one basis, into $|v\rangle$ described in terms of the other basis and vice versa.

For this particular example, $\hat{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, so

$$\hat{A}|v'\rangle = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$$

¹³also called the *image of $|e_j\rangle$ under the action of \hat{A}*

We applied \hat{A} to the description of $|v\rangle$ in terms of the new basis $\hat{A}|e_1\rangle$ and $\hat{A}|e_2\rangle$, which we labelled $|v'\rangle$, and it returned $|v\rangle$ in terms of the old, standard basis. This suggests that the matrix \hat{T}^{-1} that “returns” the coordinates of $|v\rangle$ to the old basis is precisely the matrix \hat{A} . One would just have to make an important, conceptual distinction between \hat{A} as an *active* transformation that takes $|e_1\rangle$ and $|e_2\rangle$ into a new pair of vectors $\hat{A}|e_1\rangle$ and $\hat{A}|e_2\rangle$, and \hat{A} as a *passive* transformation that takes the coordinates of $|v'\rangle$ in terms of the basis $\hat{A}|e_1\rangle$ and $\hat{A}|e_2\rangle$, and translates them into their original description in terms of $|e_1\rangle$ and $|e_2\rangle$.

Analogously, the matrix

$$\hat{A}^{-1}|v\rangle = \begin{pmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{pmatrix} \begin{pmatrix} 3 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

We applied \hat{A}^{-1} to the description of $|v\rangle$ in terms of the old, standard basis $|e_1\rangle$ and $|e_2\rangle$, and it translated $|v\rangle$ to the new basis $\hat{A}|e_1\rangle$ and $\hat{A}|e_2\rangle$. This suggests that the matrix \hat{T} that translates the coordinates of $|v\rangle$ from the old to the new basis is precisely the matrix \hat{A}^{-1} .

Notice also that

$$\hat{A}\hat{A}^{-1} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{pmatrix} = \begin{pmatrix} 4/3 - 1/3 & -2/3 + 2/3 \\ 2/3 - 2/3 & -1/3 + 4/3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

i.e.

$$\hat{A}\hat{A}^{-1} = \mathbb{1}$$

which justifies the fact that we called \hat{A}^{-1} *the inverse of \hat{A}* . In summary, the matrix \hat{A} , when seen as a passive transformation, i.e. as a change of coordinates \hat{T} , satisfies the following relations¹⁴:

$$\hat{T} = \hat{A}^{-1} \quad \& \quad \hat{T}^{-1} = \hat{A}$$

An arbitrary vector $|x\rangle$ has length $\|\vec{x}\| = \sqrt{\langle x|x\rangle}$, but it is said to be *normalised* if it is divided by its norm, thus having a new length of 1, i.e.

$$\frac{|x\rangle}{\sqrt{\langle x|x\rangle}} \quad \text{has length 1}$$

We recall that $\sqrt{\langle x|x\rangle} = \sqrt{\sum_{j=1}^n x_j^2}$, which involves the sum of its coordinates squared, but a change in coordinates will result in a different value for such a sum, and thus a different norm. This cannot be right, for any change of coordinates is merely a change of reference frame that should not induce a different value for norms of vectors. E.g.

$$\langle v|v\rangle = (3 \quad 3) \begin{pmatrix} 3 \\ 3 \end{pmatrix} = 3^2 + 3^2 = 18$$

¹⁴For an excellent course on linear algebra, see FRIEDBERG, S. H., INSEL, A. J., & SPENCE, L. E. (1989). *Linear algebra*. Englewood Cliffs, N.J., Prentice Hall.

when described with the canonical set of basis vectors, but also,

$$\langle v'|v' \rangle = (1 \quad 1) \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 1^2 + 1^2 = 2$$

when described with the new set of basis vectors. What went wrong?

We see that $\{|e_j\rangle\}_{j=1}^2$ is a set of normal, perpendicular vectors. Any set of such vectors can work as a basis for \mathbb{R}^2 , and the inner product will remain unchanged, regardless of the reference frame one chooses to describe a vector with¹⁵. So we know for sure that $\langle v|v \rangle = 18$ is the correct value. The problem before was that $\{\hat{A}|e_j\rangle\}_{j=1}^2$ is *not* a set of normal, orthogonal vectors, hence the problem to compute the right value for $\langle x|x \rangle$.

To compensate for this non-perpendicular change of coordinates $\hat{T} = \hat{A}^{-1}$, we need first to re-describe $|v'\rangle$ in terms of the original coordinate system, compute the norm, and then return to the new coordinates given in terms of $\{\hat{A}|e_j\rangle\}_{j=1}^2$. I.e. suppose we have the coordinates $|v'\rangle = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ in terms of the new frame of reference, then

$$|v'\rangle = \hat{A}^{-1}|v\rangle \quad \text{as was seen before, and}$$

$$|v\rangle = \hat{A}|v'\rangle$$

So,

$$\langle v|v \rangle = (\langle v'| \hat{A}^t) \hat{A}|v'\rangle = \langle v'| \hat{A}^t \hat{A}|v'\rangle$$

which means we can re-define a new inner product in the new frame of reference, where vectors are described in terms of the new coordinates, given by $\hat{A}|e_1\rangle$ and $\hat{A}|e_2\rangle$, as follows:

$$\langle x'|y' \rangle_A = \langle x'| \hat{C}|y' \rangle \quad \text{where}$$

$\hat{C} := \hat{A}^t \hat{A}$ is the adjustment due to the coordinate transformation

In this particular case,

$$\hat{C} = \hat{A}^t \hat{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}$$

¹⁵As we saw in §2.7, an orthogonal transformation \hat{O} is such that $\langle x|x \rangle = \langle x|\hat{O}^t \hat{O}|x \rangle$ because $\hat{O}^t \hat{O} = \mathbf{1}$. As a consequence, the computation of the inner product is invariant under orthogonal changes of basis.

Then,

$$\langle v'|v'\rangle_A = \langle v'|\hat{C}|v'\rangle = (1 \quad 1) \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 18$$

as was expected.

§2.9 A Reformulation of Known Analytical Geometry

Dirac's notation was introduced as a way to describe quantum states as vectors of a special kind of vector spaces. Its ability to express inner products and linear transformations, regardless of the complexity or dimension of the vectors spaces it deals with, in such a clean, succinct, and straightforward presentation, made it quite attractive for both physicists and mathematicians at the time. It is nowadays accepted as the standard notation for quantum mechanics, mainly because of the simplicity to perform certain computations with Dirac's notation, as opposed to any other.

To highlight the convenience and advantages of this notation, it is useful to reformulate some well known geometrical facts in terms of *bras* and *kets*. We begin by studying the cases of \mathbb{R}^2 and \mathbb{R}^3 . Any equation involving two real variables x and y has an associated set of points $\begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2$ that satisfy this equation; *satisfying* the equation means that a point in this set is such that its coordinates will always obey the relation established by this equation. E.g. the point $\begin{pmatrix} 2 \\ 4 \end{pmatrix}$ satisfies the equation $y = x^2$ because $4 = 2^2$.

So a *one-dimensional surface* in \mathbb{R}^2 , like the parabola $y = x^2$, has an associated equation involving the two variables x and y ; analogously, a two-dimensional surface in \mathbb{R}^2 has an associated equation involving the three variables x , y , and z , and this idea can be generalised to any dimension. The different properties of these surfaces depends on the relation between the different variables, and not all surfaces are smooth and continuous, but we can focus on a few particular examples to grasp the idea fully before trying to understand the *pathological* cases.

Circles, Spheres, and Hyperspheres

The equation of a circle in \mathbb{R}^2 is given, by definition, by the points in the plane that lie equidistant to the centre, the origin for example. So for a fixed length r , any point in the plane at this distance from the origin lies in the circle of radius r ; the corresponding equation is

$$d_{[\vec{0}, \vec{x}]} = x^2 + y^2 = r$$

Or, in the new notation

$$\langle x|x\rangle = (x \quad y) \begin{pmatrix} x \\ y \end{pmatrix} = x^2 + y^2 = r^2$$

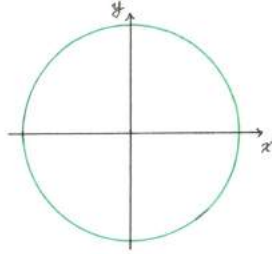


Figure 2.14: A circle $C \subset \mathbb{R}^2$

which is only a different way of saying that the distance between the point $|x\rangle = \begin{pmatrix} x \\ y \end{pmatrix}$ and the origin is precisely r . If one applies this definition to points $|x\rangle \in \mathbb{R}^3$, the resulting set of points is the surface of a sphere. Of course, a sphere of radius r , centred at the origin in \mathbb{R}^3 , has an equation

$$\langle x|x\rangle = (x \quad y \quad z) \begin{pmatrix} x \\ y \\ z \end{pmatrix} = x^2 + y^2 + z^2 = r^2$$

The hypersphere of radius r , centred at the origin in \mathbb{R}^4 , has an equation

$$\langle x|x\rangle = (x \quad y \quad z \quad w) \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = x^2 + y^2 + z^2 + w^2 = r^2$$

and any arbitrary, n -dimensional hypersphere of radius r , centred at the origin in \mathbb{R}^n , has an equation

$$\langle x|x\rangle = (x_1 \quad x_2 \quad \dots \quad x_n) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = r^2$$

So, in general, the equation $\langle x|x\rangle = r^2$ suggests the idea of a sphere, a fixed length from the origin, and the dimension of this sphere is left out of the equation.

Ellipses

Given two fixed points, F_1 and F_2 in \mathbb{R}^2 , and a fixed distance d , the set of points whose distance to F_1 plus their distance to F_2 add up to the distance d , lie in

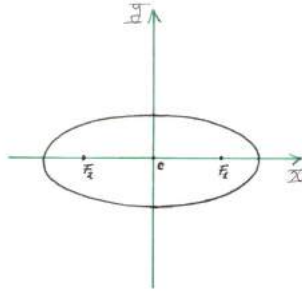


Figure 2.15: An ellipse $E \subset \mathbb{R}^2$

an ellipse with foci F_1 and F_2 . They satisfy the following relation¹⁶:

$$\alpha x^2 + \beta y^2 = c$$

where α and β are two real, positive parameters. Or, in the new notation

$$\langle x | \hat{E} | x \rangle = (x \ y) \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \alpha x^2 + \beta y^2 = c$$

Hyperbolae

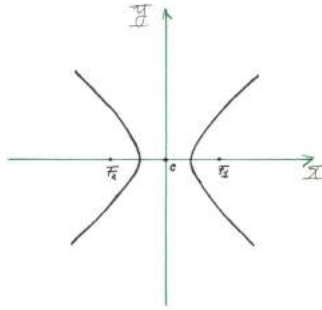


Figure 2.16: Hyperbola $H \subset \mathbb{R}^2$

Given two fixed points, F_1 and F_2 in \mathbb{R}^2 , and a fixed distance d , the set of points whose distance to F_1 minus their distance to F_2 is exactly d lie in an

¹⁶which can be easily obtained from the canonical form: $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ by defining $a := \sqrt{\frac{-c}{\alpha}}$ and $b := \sqrt{\frac{-c}{\beta}}$. It should be clear that the constant $c < 0$.

hyperbola with foci F_1 and F_2 . They satisfy the following relation¹⁷:

$$\alpha x^2 - \beta y^2 = c \quad (2.1)$$

where α and β are two real, positive parameters. Or, in the new notation

$$\langle x | \hat{H} | x \rangle = (x \ y) \begin{pmatrix} \alpha & 0 \\ 0 & -\beta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \alpha x^2 - \beta y^2 = c$$

Parabolas

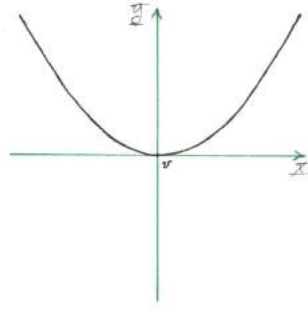


Figure 2.17: Parabola $P \subset \mathbb{R}^2$

Given a fixed point, F_1 in \mathbb{R}^2 , and a fixed line L , the set of points whose distance to F_1 equals their distance to the line L lie in a parabola with focus F_1 . They satisfy the following relation¹⁸:

$$\alpha x^2 + ey = c$$

where α and e are two real parameters. Or, in the new notation

$$\langle x | \hat{P} | x \rangle + \langle d | x \rangle = (x \ y) \begin{pmatrix} \alpha & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + (0 \ e) \begin{pmatrix} x \\ y \end{pmatrix} = \alpha x^2 + ey = c$$

where $|d\rangle = \begin{pmatrix} 0 \\ e \end{pmatrix}$. So, in general, any polynomial equation describing a conical section $\alpha x^2 + \beta y^2 + \gamma xy + dx + ey + k = 0$, that can be displaced from the origin, thus the linear part $dx + ey$, and/or rotated by an angle θ , thus the non-zero term γxy , can be expressed in Dirac's notation as $\langle x | \hat{A} | x \rangle + \langle d | x \rangle + c = 0$, since

$$\langle x | \hat{A} | x \rangle + \langle d | x \rangle + c = (x \ y) \begin{pmatrix} \alpha & \gamma/2 \\ \gamma/2 & \beta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + (d \ e) \begin{pmatrix} x \\ y \end{pmatrix} + c$$

¹⁷which, again, can be easily obtained from the canonical form: $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$ by defining $a := \sqrt{\frac{-c}{\alpha}}$ and $b := \sqrt{\frac{-c}{\beta}}$. It should be clear that the constant $c < 0$.

¹⁸usually expressed in the canonical form $y = ax^2 + b$

$$= \alpha x^2 + \gamma xy + \beta y^2 + dx + ey + c$$

and we can classify conic sections in terms of the properties of their associated linear transformations as follows:

Circle $\langle x|x\rangle = \langle x|\mathbf{1}|x\rangle = r^2$

$$\mathbf{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \implies \det \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = 1$$

Ellipse $\langle x|\hat{E}|x\rangle = c$

$$\hat{E} = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix} \implies \det \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix} = \alpha\beta > 0$$

Hyperbola $\langle x|\hat{H}|x\rangle = c$

$$\hat{H} = \begin{pmatrix} \alpha & 0 \\ 0 & -\beta \end{pmatrix} \implies \det \begin{pmatrix} \alpha & 0 \\ 0 & -\beta \end{pmatrix} = -\alpha\beta < 0$$

Parabola $\langle x|\hat{P}|x\rangle + \langle d|y\rangle = c$

$$\hat{P} = \begin{pmatrix} \alpha & 0 \\ 0 & 0 \end{pmatrix} \quad \& \quad |d\rangle = \begin{pmatrix} 0 \\ e \end{pmatrix} \implies \det \begin{pmatrix} \alpha & 0 \\ 0 & 0 \end{pmatrix} = 0$$

§2.10 Complex Numbers

Complex numbers are frequently used in physics. They are useful, and almost indispensable, to describe wave functions, light polarisation and even the most elementary quantum systems, so we now take a moment to study their origin, geometry and some algebraic properties.

Anyone who has tried to solve a quadratic equation of the form $ax^2 + bx + c = 0$ knows that a problem arises when $(b^2 - 4ac)$ in $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ is a negative number, since square roots only exist for positive numbers. The problem can be solved by defining the quantity $i := \sqrt{-1}$ and providing any quadratic equation with a general solution of the form $a + ib$, where a and b are real numbers¹⁹.

The geometrical properties of this new element i are not often discussed. Let us first see what the operation “*multiply by -1*” means geometrically. We depict the set of real numbers by a straight line that extends from $-\infty$ to ∞ centred at 0. A positive, real number a lies on this line, to the right of 0, as

¹⁹A more adequate definition for i is given by $i^2 := -1$, to avoid the possible confusion with $i^2 = i \cdot i = \sqrt{-1}\sqrt{-1} = \sqrt{(-1)(-1)} = \sqrt{1} = 1$

shown in Figure 2.18. Multiplying a by -1 results in a 180° or π rotation from the origin.

This suggests that the operation “multiply by i ” is such, that performed twice results in a multiplication by -1 . This is equivalent to a π rotation, so $a \cdot i$ would be geometrically equivalent to a 90° or $\frac{\pi}{2}$ rotation. Thus, multiplication by i done twice is equivalent to multiplication by i^2 , i.e. -1 .

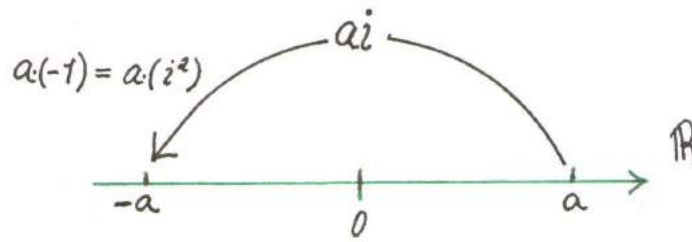


Figure 2.18: Geometric meaning of *multiplication by i* and i^2 .

Figure 2.18 illustrates this relation, which is a nice depiction of i and the way it interacts with the set of real numbers; it clearly does not belong in \mathbb{R} , but it seems to be “floating” on top of it. If i lies *over* the line of real numbers, then so do $2i$, $3i$, $-i$, $-2i$, and all other scalar factors of i . This also suggests a very clear identification between the set of complex numbers \mathbb{C} (numbers of the form $x + iy$), and the plane \mathbb{R}^2 , as is shown in Figure 2.19.

One should be careful though, when comparing the plane \mathbb{R}^2 and the set \mathbb{C} ; the former has two real dimensions, but the latter could be said to have just one complex, or “imaginary” dimension, since one needs only one (complex) number to describe it fully. \mathbb{R}^2 and \mathbb{R}^3 are said to be two, and three-dimensional, respectively, because one needs two and three real numbers to refer to any of its elements. Any complex number z , however, can be uniquely and unequivocally referred to with just one complex number, itself ²⁰.

One of the most interesting features of the plane \mathbb{C} , as opposed to \mathbb{R}^2 , is its algebraic structure. Vectors $|x\rangle \in \mathbb{R}^2$ can be added and multiplied by scalars, but they cannot be multiplied by other vectors. Complex numbers, however, like real numbers, can be added and multiplied. Let us see how this works algebraically. If $z = a + ib$ and $w = c + id$, then

$$z \cdot w = (a + ib)(c + id) = ac + i^2bd + iad + ibc = (ac - bd) + i(ad + bc)$$

What would this multiplication look like, provided we identify z and w with

²⁰One could argue that z is composed of two real numbers, namely $z = x + iy$, and this is true, so \mathbb{C} has one complex dimension, but two real dimensions. This is why it is isomorphic to \mathbb{R}^2 .

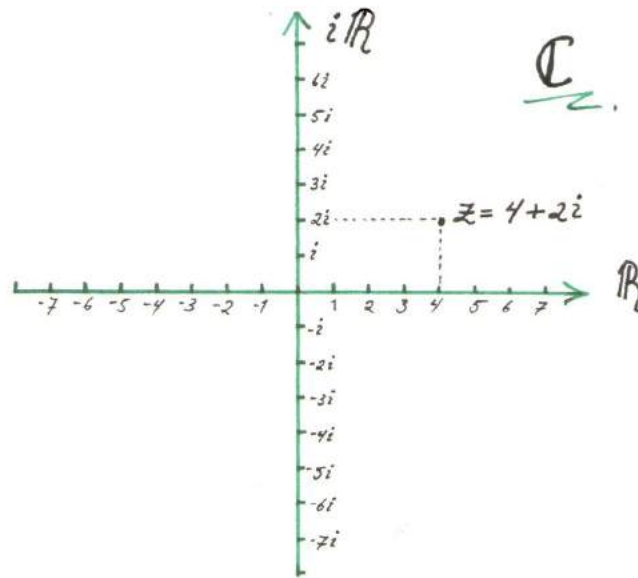


Figure 2.19: The plane of complex numbers \mathbb{C} .

their corresponding vectors in \mathbb{R}^2 ? To see this, we first identify

$$z \in \mathbb{C} \quad \text{where} \quad z = a + ib \quad \text{to} \quad \begin{pmatrix} a \\ b \end{pmatrix} \in \mathbb{R}^2$$

as was done for z in Figure 2.19, and

$$w \in \mathbb{C} \quad \text{where} \quad w = c + id \quad \text{to} \quad \begin{pmatrix} c \\ d \end{pmatrix} \in \mathbb{R}^2$$

This way, we can perform the same identification process, and interpret the result of zw geometrically as the new complex number

$$z \cdot w = (ac - bd) + i(ad + bc) \quad \implies \quad \begin{pmatrix} ac - bd \\ ad + bc \end{pmatrix}$$

This establishes the main difference between the plane \mathbb{R}^2 and the complex plane \mathbb{C} . The *algebra* of complex numbers allows not just addition (the equivalent of vector addition if we identify complex numbers with vectors in \mathbb{R}^2) but also multiplication²¹. This multiplication by a single complex number can be seen as a matrix multiplication provided we identify one of the complex numbers to a matrix as follows:

$$z \cdot w = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} ac - bd \\ ad + bc \end{pmatrix}$$

²¹not just scalar multiplication.

and since complex number multiplication commutes, this is equivalent to

$$w \cdot z = \begin{pmatrix} c & -d \\ d & c \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} ac - bd \\ ad + bc \end{pmatrix}$$

As an example, $z = 5 + i$ and $w = 1 + 2i$ in Figure 2.20 can be multiplied to obtain a new complex number $z \cdot w$ whose length $\|z \cdot w\| = \|z\| \cdot \|w\|$, and whose angle with respect to the real axis is the sum of the angles z and w make with the real axis.

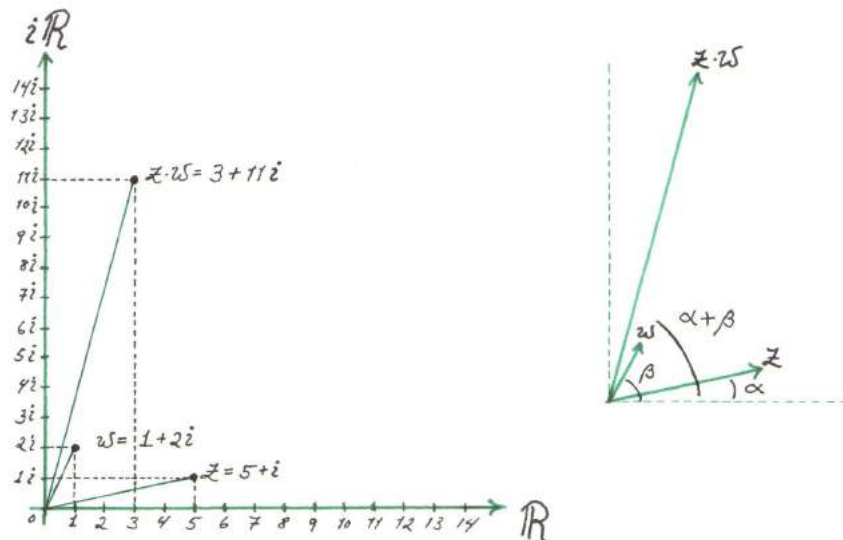


Figure 2.20: Matrix multiplication of complex numbers and its geometrical interpretation.

The norm of any complex number $z = a + ib$ has to correspond to the norm of its associated vector in \mathbb{R}^2 , i.e. $\|z\| = a^2 + b^2$, so we define a new complex number $z^* = a - ib$ so that

$$\|z\|^2 = z^* \cdot z = (a - ib)(a + ib) = a^2 + b^2$$

A single complex number z can be decomposed as a pair of 2×2 matrices to determine completely how the algebra of complex numbers works. I.e. for every $z \in \mathbb{C}$,

$$z = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} + \begin{pmatrix} 0 & -b \\ b & 0 \end{pmatrix}$$

or equivalently ²²,

$$z = a + ib = a\mathbf{1} + b\hat{i}$$

²²One should check how this is compatible with complex number multiplication as defined previously, i.e.

$$\begin{aligned} z \cdot w &= \left[\begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} + \begin{pmatrix} 0 & -b \\ b & 0 \end{pmatrix} \right] \left[\begin{pmatrix} c & 0 \\ 0 & c \end{pmatrix} + \begin{pmatrix} 0 & -d \\ d & 0 \end{pmatrix} \right] \\ &= \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} \begin{pmatrix} c & 0 \\ 0 & c \end{pmatrix} + \begin{pmatrix} 0 & -b \\ b & 0 \end{pmatrix} \begin{pmatrix} c & 0 \\ 0 & c \end{pmatrix} + \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} \begin{pmatrix} 0 & -d \\ d & 0 \end{pmatrix} + \begin{pmatrix} 0 & -b \\ b & 0 \end{pmatrix} \begin{pmatrix} 0 & -d \\ d & 0 \end{pmatrix} \\ &= \begin{pmatrix} ac & 0 \\ 0 & ac \end{pmatrix} + \begin{pmatrix} 0 & -bc \\ bc & 0 \end{pmatrix} + \begin{pmatrix} 0 & -ad \\ ad & 0 \end{pmatrix} + \begin{pmatrix} -bd & 0 \\ 0 & bd \end{pmatrix} \\ &= \begin{pmatrix} ac - bd & 0 \\ 0 & ac - bd \end{pmatrix} + \begin{pmatrix} 0 & -(bc + ad) \\ bc + ad & 0 \end{pmatrix} = z \cdot w \end{aligned}$$

if we define

$$\mathbf{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \& \quad \hat{i} := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

If we wish to create complex spaces of greater dimension as we do with \mathbb{R} , \mathbb{R}^2 , \mathbb{R}^3 , and so on, we need to redefine vector transposition to keep consistency in the way we compute norms. So for $|z\rangle \in \mathbb{C}^2$, which is isomorphic to \mathbb{R}^4 since it has four different parameters,

$$|z\rangle = \begin{pmatrix} z \\ w \end{pmatrix} \quad \text{where} \quad z, w \in \mathbb{C}$$

This means that

$$\langle z| = (z^* \quad w^*)$$

so that

$$\langle z|z\rangle = (z^* \quad w^*) \begin{pmatrix} z \\ w \end{pmatrix} = z \cdot z^* + w \cdot w^* = a^2 + b^2 + c^2 + d^2$$

It is easy to identify how these two-dimensional numbers \mathbb{C} are closely related to matrices, and thus how they differ from regular, one-dimensional numbers (real numbers). The Irish mathematician William R. Hamilton discovered, whilst going for a walk with his wife, that there are no three-dimensional numbers; the next possible algebra is that of four-dimensional numbers, called *quaternions*. From these, only eight-dimensional numbers (*octonions*), sixteen-dimensional numbers (*sedonions*), etc. can exist. Each step up means a loss of an algebraic property, e.g. complex numbers have no order, meaning there is no such thing as $z > w$, quaternion multiplication is not commutative, octonion multiplication is neither commutative nor associative, and so on.

§2.11 Matrices: Some Examples of Different Interpretations

Besides linear transformations of \mathbb{R}^2 and \mathbb{R}^3 , matrices can be interpreted and used in a wide variety of situations. One of the most common ways to introduce matrices is with systems of linear equations. These systems have also a geometric approach, which might seem more familiar provided one has previously worked with matrices as linear transformations. We begin by studying systems of linear equations with two variables.

Systems of Linear Equations

A line L is a subset of \mathbb{R}^2 with an associated equation $ax + by = c$. The point of intersection of any two lines $L_1 : a_1x + b_1y = c_1$ and $L_2 : a_2x + b_2y = c_2$ in the plane has coordinates $\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ that satisfy both equations, i.e.

$$\begin{aligned} a_1x_0 + b_1y_0 &= c_1 \\ a_2x_0 + b_2y_0 &= c_2 \end{aligned}$$

which can also be seen as a linear transformation $\hat{A} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ such that

$$\hat{A}|x\rangle = \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

where the determinant of \hat{A} plays an important role, since

$$\hat{A}|x\rangle = |c\rangle$$

If we multiply \hat{A}^{-1} from the left

$$\hat{A}^{-1}\hat{A}|x\rangle = \hat{A}^{-1}|c\rangle$$

$$\mathbb{1}|x\rangle = \hat{A}^{-1}|c\rangle$$

$$|x\rangle = \hat{A}^{-1}|c\rangle$$

but

$$\hat{A}^{-1} = \frac{1}{\det \hat{A}} \begin{pmatrix} b_2 & -b_1 \\ -a_2 & a_1 \end{pmatrix}$$

so the system $\hat{A}|x\rangle = |c\rangle$ has a solution if and only if $\det(\hat{A}) \neq 0$. Otherwise, the two lines L_1 and L_2 are either parallel or the same line.

This result can be extended to any system of n variables, where

$$\hat{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}, \quad |x\rangle = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \& \quad |c\rangle = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}$$

and, again, the system $\hat{A}|x\rangle = |c\rangle$ has a solution if and only if $\det(\hat{A}) \neq 0$.

System of Linear Differential Equations

Suppose the position in space \vec{x} of a particle is given as a function of time as

$$\vec{x}(t) = \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}. \quad \text{Then the velocity vector at a certain time } t_0 \text{ corresponds to}$$

the time derivative of this vector, i.e.

$$\frac{d\vec{x}}{dt} = \begin{pmatrix} \frac{dx(t)}{dt} \\ \frac{dy(t)}{dt} \\ \frac{dz(t)}{dt} \end{pmatrix}$$

If each individual velocity is a linear combination of the coordinates as functions of time, this system of differential equations can be expressed as follows:

$$\frac{dx}{dt} = a_1x(t) + b_1y(t) + c_1z(t)$$

$$\frac{dy}{dt} = a_2x(t) + b_2y(t) + c_2z(t)$$

$$\frac{dz}{dt} = a_3x(t) + b_3y(t) + c_3z(t)$$

and thus, the velocity vector $\frac{d\vec{x}}{dt}$ can be expressed as a linear transformation as

$$\begin{pmatrix} \frac{dx(t)}{dt} \\ \frac{dy(t)}{dt} \\ \frac{dz(t)}{dt} \end{pmatrix} = \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix} \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}$$

and accordingly, the system $\vec{v} = \hat{A}\vec{x}$ has a solution if and only if $\det(\hat{A}) \neq 0$.

N^{th} -order Homogeneous Differential Equation

Any homogeneous differential equation of the form²³ $\frac{d^3y}{dt^3} + a \cdot \frac{d^2y}{dt^2} + b \cdot \frac{dy}{dt} + c \cdot y(t) = 0$ can be reduced to a system of linear equations of degree 1 by defining the following variables:

$$x_1 := y(t)$$

$$x_2 := \frac{dy}{dt}$$

$$x_3 := \frac{d^2y}{dt^2}$$

This transforms the previous equation into $\frac{dx_3}{dt} + ax_3 + bx_2 + cx_1 = 0$ or,

²³An equation of the form $a \cdot \frac{d^3y}{dt^3} + b \cdot \frac{d^2y}{dt^2} + c \cdot \frac{dy}{dt} + d \cdot y(t) = 0$ can easily be transformed into an equation like the one above by dividing everything by a , i.e. $\frac{d^3y}{dt^3} + \frac{b}{a} \cdot \frac{d^2y}{dt^2} + \frac{c}{a} \cdot \frac{dy}{dt} + \frac{d}{a} y(t) = 0$

equivalently

$$\begin{pmatrix} \frac{dx_3}{dt} \\ \frac{dx_2}{dt} \\ \frac{dx_1}{dt} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -c & -b & -a \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

which can be generalised for homogeneous equations of any order n .

§2.12 Dual Spaces

A linear function in general may not only transform vectors from one vector space V into itself, but also between different vector spaces. A linear transformation $\hat{T} : V \rightarrow W$ may transform, for example, \mathbb{R}^2 into \mathbb{R}^3 , and one would represent \hat{T} with a 3×2 matrix. Analogously, if \hat{T} transforms \mathbb{R}^3 into \mathbb{R}^2 , one would then represent \hat{T} with a 2×3 matrix.

\mathbb{R} , \mathbb{R}^2 , \mathbb{R}^3 , \dots , \mathbb{R}^n , \dots are all canonical examples of vector spaces with elements $|x\rangle$, but the set of linear transformations between two vector spaces is also a vector space, e.g. $M_{2 \times 2}(\mathbb{R})$ is a vector space with elements of the form

$$\hat{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

In particular, the set of linear transformations $\langle x| : V \rightarrow \mathbb{R}$ is also a vector space, since it takes a vector $|x\rangle \in V$ and transforms it into a scalar $a \in \mathbb{R}$ via left matrix multiplication. I.e. if the space formed by $|x\rangle$ is a vector space (as \mathbb{R}^2 or \mathbb{R}^3), then the space of $\langle x|$ is also a vector space, consisting of $1 \times n$ matrices that form a set of linear transformations from \mathbb{R}^n to \mathbb{R} .

For any arbitrary vector space V , this new vector space is denoted as V^* and is called the *dual space*; it consists of all linear transformations from V to the field of real numbers²⁴ \mathbb{R} . Since every $|x\rangle \in V$ has a dual element $\langle x| \in V^*$, the dimension of V^* is exactly the dimension of V . One should be careful with these last two assertions and the way one associates an element with its dual, for there are certain subtleties that are only true for finite dimensional vector spaces; it is, however, a good starting point to understand further studies in linear algebra.

The basis of V has a corresponding dual basis; for example, \mathbb{R}^3 has its associated dual space, and the new basis is obtained by transposing

$$\begin{aligned} |e_1\rangle &= \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} & \implies & \langle e_1| = (1 \quad 0 \quad 0) \\ |e_2\rangle &= \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} & \implies & \langle e_2| = (0 \quad 1 \quad 0) \end{aligned}$$

²⁴or the field of complex numbers \mathbb{C} in case V is defined as a complex vector space

$$|e_3\rangle = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \implies \langle e_3| = (0 \ 0 \ 1)$$

This implies that a vector, decomposed as a linear combination of the basis elements, has an associated dual vector given by

$$|x\rangle = x|e_1\rangle + y|e_2\rangle + z|e_3\rangle \implies \langle x| = x\langle e_1| + y\langle e_2| + z\langle e_3|$$

for $x, y, z \in \mathbb{R}$ the vector's coordinates²⁵.

Notice that

$$\langle e_1|e_1\rangle = (1 \ 0 \ 0) \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = 1 \quad , \quad \langle e_1|e_2\rangle = (1 \ 0 \ 0) \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = 0$$

$$\langle e_1|e_3\rangle = (1 \ 0 \ 0) \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = 0$$

Analogously,

$$\langle e_2|e_1\rangle = 0 \quad , \quad \langle e_2|e_2\rangle = 1 \quad \& \quad \langle e_2|e_3\rangle = 0$$

$$\langle e_3|e_1\rangle = 0 \quad , \quad \langle e_3|e_2\rangle = 0 \quad \& \quad \langle e_3|e_3\rangle = 1$$

which is often abbreviated as $\langle e_i|e_j\rangle = \delta_{ij}$, where

$$\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

Vectors satisfy the previous condition if and only if they are orthogonal. Vectors that are both normal and orthogonal are said to be *orthonormal*.

§2.13 Outer Products and Projectors

In §1.6 and §1.7 of this chapter we defined the operation *inner product* of a *bra* and a *ket* as the scalar $\langle x|y\rangle$. We now define the *outer product* of a *ket* and a *bra* as $|x\rangle\langle y|$ and study its geometrical properties.

²⁵or

$$|z\rangle = z|e_1\rangle + w|e_2\rangle + u|e_3\rangle \implies \langle z| = z^*\langle e_1| + w^*\langle e_2| + u^*\langle e_3|$$

in the case of a vector in a complex vector space like $|z\rangle \in \mathbb{C}^3$.

For two elements of a vector space and its dual $|y\rangle$ and $\langle x|$, the result of the outer product $|x\rangle\langle y|$ is a linear transformation \hat{O}

$$|x\rangle\langle y| = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} (y_1 \quad y_2 \quad \dots \quad y_n) = \begin{pmatrix} x_1 y_1 & x_1 y_2 & \dots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \dots & x_2 y_n \\ \vdots & & & \\ x_n y_1 & x_n y_2 & \dots & x_n y_n \end{pmatrix}$$

which can be then applied to another vector $|a\rangle \in V$. For example, $|x\rangle\langle y|$ applied as a linear function to the vector $|a\rangle$ is

$$|x\rangle\langle y|(|a\rangle) = |x\rangle\langle y|a\rangle = (\langle y|a\rangle)|x\rangle$$

means the vector $|a\rangle$ is first projected to $|y\rangle$, and then this factor scales the vector $|x\rangle$; the result is a vector pointing out in the $|x\rangle$ direction. In particular, for a vector $|x\rangle$ of length 1, we have

$$|x\rangle\langle x|(|a\rangle) = |x\rangle\langle x|a\rangle = \langle x|a\rangle|x\rangle$$

is an operator that finds the projection of $|a\rangle$ on the vector $|x\rangle$, and is then directed in the $|x\rangle$ direction, since $\|\vec{x}\| = 1 \implies \langle x|a\rangle = \|\vec{x}\| \cdot \|\vec{a}\| \cdot \cos(\theta) = \|\vec{a}\| \cdot \cos(\theta) = P_{ax}$.

So we define the *projector* of ket $|x\rangle$ as²⁶

$$\hat{P}_x = |x\rangle\langle x|$$

§2.14 Function Spaces

So far we have studied examples of finite-dimensional vector spaces. Our first example of an infinite dimensional vector space is the set of real-valued functions

$$\mathcal{F} = \{f : I \subseteq \mathbb{R} \longrightarrow \mathbb{R} \mid f \text{ is continuous}\}$$

which has a vector space structure if we identify functions $f, g \in \mathcal{F}$ as vectors with pointwise addition, i.e.

$$(f + g)(x) := f(x) + g(x) \quad \text{for all } x \in I \subseteq \mathbb{R}$$

and scalar multiplication given by

$$(\lambda f)(x) := \lambda f(x) \quad \text{for all } x \in I \subseteq \mathbb{R}$$

Once we have defined both vector addition and scalar multiplication, the algebraic structure we have provided makes it clear that it is, in fact, a vector space. This does not show, however, what this abstract space might actually look like. Moreover, finding a proper basis for this vector space means finding a subset of

²⁶Not to be confused with \hat{p}_x a momentum operator in the x -dimension.

functions $\beta \subset \mathcal{F}$ such that any other function f could be uniquely expressed as a linear combination of elements of β .

A more simple example can be useful to grasp the idea of a function vector space. Suppose we narrow the spectrum of real functions to cubic polynomials, i.e. polynomials of degree at most 3. That being the case, every function f has the form

$$f(x) = a + bx + cx^2 + dx^3$$

where $a, b, c, d \in \mathbb{R}$. We can define a basis $\beta = \{f_0(x) = 1, f_1(x) = x, f_2(x) = x^2, f_3(x) = x^3\}$ and identify any polynomial of degree 3 as a linear combination of this set of basis elements, i.e. every cubic function can be expressed in terms of these four functions. The function $f(x) = 3x^3 - 2x^2 + 1$, for example, can be decomposed as follows:

$$f(x) = 1 \cdot f_0(x) + 0 \cdot f_1(x) - 2 \cdot f_2(x) + 3 \cdot f_3(x)$$

and could be expressed in terms of its coordinates as

$$|f\rangle = \begin{pmatrix} 1 \\ 0 \\ -2 \\ 3 \end{pmatrix}$$

which clarifies how β is a basis for the vector space of cubic polynomials. This function space is generated by four different, independent polynomials, so the dimension of this vector space is 4.

Similarly, the function space of polynomials of degree at most n is an $(n+1)$ -dimensional space generated by the basis $\beta = \{f_0(x) = 1, f_1(x) = x, f_2(x) = x^2, f_3(x) = x^3, \dots, f_n(x) = x^n\}$. Thus, every polynomial function of degree n can be expressed as a linear combination of basis elements as follows:

$$f(x) = \lambda_0 \cdot f_0(x) + \lambda_1 \cdot f_1(x) + \lambda_2 \cdot f_2(x) + \lambda_3 \cdot f_3(x) + \dots + \lambda_n \cdot f_n(x)$$

i.e.

$$|f\rangle = \begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix}$$

Finally, the function space of all polynomials of arbitrary degree is an infinite dimensional vector space with basis $\beta = \{f_0(x) = 1, f_1(x) = x, f_2(x) = x^2, f_3(x) = x^3, \dots, f_n(x) = x^n, \dots\}$ and every element has an infinite array of coordinates corresponding to the series of coefficients that constitute its expansion as a linear combination of basis functions.

§2.15 Taylor Series

The way to describe a polynomial as a vector, i.e. a linear combination of single functions of degree 0 through n , is quite straightforward. A more general function like $f(x) = \sin(x)$, however, does not seem to have any alternative to use this kind of approach. Any polynomial function that approached $\sin(x)$ at any given point would merely approximate it within a region $(a, b) \subset \mathbb{R}$.

For a given x close to $x = 0$, the linear function $f_1(x) = x$ seems like a good approximation of $\sin(x)$; once we get closer to $x = \frac{\pi}{2}$, these two functions diverge significantly. By adding a cubic term, $f(x) = -\frac{x^3}{3!}$ we obtain a new function that resembles $\sin(x)$ in a larger region around $x = 0$, but again, these two functions diverge once x is *far* from the range $(-\pi, \pi)$. Figure 2.21 shows how polynomials can approach the function $f(x) = \sin(x)$ as we add terms of higher order.

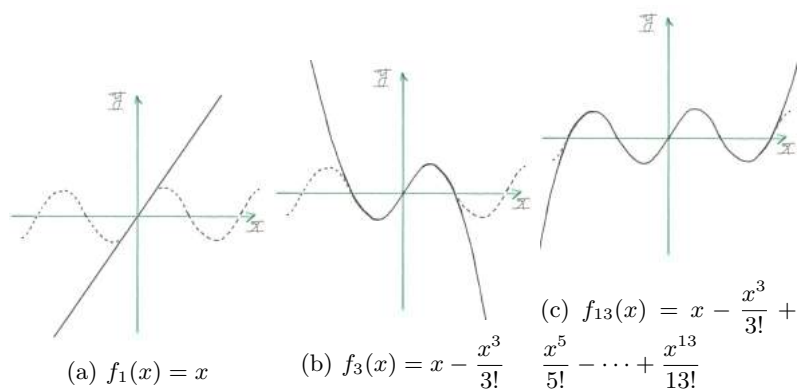


Figure 2.21: Polynomials can approach the function $f(x) = \sin(x)$ as we add terms of higher order.

The coefficients we used to approach this function with polynomials were obtained with the following reasoning. We want, ideally, to express the function $f(x) = \sin(x)$ as a linear combination of powers of x . Suppose we will use a polynomial of degree n , labelled $P_n(x)$.

$$P_n(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n$$

For a particular value, e.g. $x = 0$, we need $f(0) = P_n(0)$ if $P_n(x)$ is to approach $\sin(x)$. If we want $P_n(x)$ to match $f(x)$ verbatim, we will need more than one point convergence; after all, there are infinitely many functions that coincide with $\sin(x)$ at $x = 0$. A stronger condition would be for $P_n(x)$ to have the same slope at $x = 0$ than $f(x) = \sin(x)$. This translates into the following equation at $x = 0$:

$$\frac{dP_n}{dx} = \frac{df}{dx}$$

Accordingly, we can extend this logic and ask, as a condition, that *every* derivative of $P_n(x)$ coincides with the corresponding derivative of $f(x)$ at $x = 0$. I.e.

$$\begin{aligned} P_n(x) &= f(x) \\ \frac{dP_n}{dx} &= \frac{df}{dx} \\ \frac{d^2P_n}{dx^2} &= \frac{d^2f}{dx^2} \\ &\vdots \\ \frac{d^n P_n}{dx^n} &= \frac{d^n f}{dx^n} \end{aligned}$$

If all these conditions are met simultaneously as $n \rightarrow \infty$, the two functions could actually be recognised as *identical*. I.e. if we solve for the coefficients a_j in the following system,

$$\begin{aligned} a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n &= \sin(x) \\ a_1 + 2a_2x + 3a_3x^2 + \cdots + na_nx^{n-1} &= \cos(x) \\ 2a_2 + 3 \cdot 2a_3x + \cdots + n \cdot (n-1)a_nx^{n-2} &= -\sin(x) \\ &\vdots \\ n!a_n &= \frac{d^n f}{dx^n} \end{aligned}$$

we can then evaluate these last $n + 1$ equations in the desired value, e.g. $x = 0$, and find these coefficients. The resulting polynomial $P_n(x)$ is called the *Taylor expansion*²⁷ or *Taylor series* of $f(x)$. It provides a unique decomposition of continuous, smooth real functions in terms of the coefficients a_j .

²⁷Maclaurin expansion if it is specifically around $x = 0$

§2.16 Linear Combinations and Infinite Basis

A function like $f(x) = \sin(x)$ has a Taylor expansion as power series that corresponds to its coordinates when seen as a vector $|f\rangle \in \mathcal{F}$. It is advisable to know and grow acquainted with the most common Taylor expansions. They are useful to solve differential equations when the exact solution cannot be found numerically, since they provide an approximation that can be as precise as needed depending on the degree one chooses to work with. To have a better approximation, one just takes more terms of the power series into account.

Frequently, taking the approximation of degree 2 is more than enough to obtain a satisfactory solution. It is important to notice that a function f is *only* equal to its Taylor expansion if one considers the whole, infinite sum. This sum must be equal to the function when evaluated at a particular value x , so the issue of convergence is non-trivial. One has to check that a Taylor series does not diverge to infinity at certain values. The most common expansions are described below.

The function $f(x) = \sin(x)$ is expanded as an alternating power series as

$$\sin(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} \dots \quad \text{where} \quad \begin{pmatrix} 0 \\ 1 \\ 0 \\ -\frac{1}{3!} \\ 0 \\ \frac{1}{5!} \\ \vdots \end{pmatrix}$$

would be the infinite set of coordinates when described as a vector $|f\rangle \in \mathcal{F}$, in terms of the basis $\beta = \{f_0(x) = 1, f_1(x) = x, f_2(x) = x^2, f_3(x) = x^3, \dots, f_n(x) = x^n, \dots\}$, defined in §2.14. This suggests why $\sin(x)$ is commonly said to be an *odd* function.

The function $f(x) = \cos(x)$ is expanded as an alternating power series as

$$\cos(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} \dots \quad \text{where} \quad \begin{pmatrix} 1 \\ 0 \\ -\frac{x^2}{2!} \\ 0 \\ \frac{x^4}{4!} \\ \vdots \end{pmatrix}$$

would be the infinite set of coordinates when described as a vector $|f\rangle \in \mathcal{F}$, in terms of the basis β . This suggests why $\cos(x)$ is commonly said to be an *even* function.

The function $f(x) = e^x$ is expanded as a power series as

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!} \dots \quad \text{where} \quad \begin{pmatrix} 1 \\ 1 \\ \frac{1}{2!} \\ \frac{1}{3!} \\ \vdots \end{pmatrix}$$

would be the infinite set of coordinates when described as a vector $|f\rangle \in \mathcal{F}$, in terms of the basis β . Differentiating this function as a polynomial makes it clear why $\frac{d}{dx}e^x = e^x$

As was discussed before, one needs, in general, an infinite amount of coefficients to describe a smooth, real function $f(x)$ as an element of the function space \mathcal{F} .

§2.17 Inner Product

In previous sections we defined the inner product of two vectors as $\langle x|y\rangle = \sum_{j=1}^n x_j y_j$ where x_j is the j^{th} coordinate of $|x\rangle$. If we have two functions f, g as vectors (and we shall now represent them as *kets*) described by the sequences

$$|f\rangle = \begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \lambda_2 \\ \vdots \end{pmatrix} \quad \text{and} \quad |g\rangle = \begin{pmatrix} \xi_0 \\ \xi_1 \\ \xi_2 \\ \vdots \end{pmatrix}$$

then their inner product should be given by the infinite sum

$$\langle f|g\rangle = \sum_{j=1}^{\infty} \lambda_j \xi_j$$

where λ_j and ξ_j are the functions' Fourier coefficients. If one wants to work with complex-valued functions, also expressed in their Fourier coefficients, then their inner product is given by the following relation:

$$\langle f|g\rangle = (\lambda_0^* \quad \lambda_1^* \quad \lambda_2^* \quad \dots) \begin{pmatrix} \xi_0 \\ \xi_1 \\ \xi_2 \\ \vdots \end{pmatrix} = \sum_{j=1}^{\infty} \lambda_j^* \xi_j$$

The reasons for this were discussed in §2.10 *Complex Numbers*.

As was discussed in §2.8, this computational process only works when we describe vectors in terms of an orthonormal basis. The basis β used in Taylor expansions is *not* an orthonormal basis, but there is another particular set of functions that can be used as an orthonormal basis for the vector space of real-valued functions.

The set of *Fourier functions*, given by

$$\begin{aligned} f_1(x) &= \sin\left(\frac{2\pi}{\lambda}x\right) \\ f_2(x) &= \sin\left(\frac{2\pi}{\lambda}2 \cdot x\right) \\ f_3(x) &= \sin\left(\frac{2\pi}{\lambda}3 \cdot x\right) \\ &\vdots \\ f_n(x) &= \sin\left(\frac{2\pi}{\lambda}n \cdot x\right) \end{aligned}$$

is in fact an orthonormal basis of \mathcal{F} . This means that *every* function $f(x)$ can be expressed as a series of harmonic functions²⁸. The coefficients of $f(x)$ obtained from its Fourier expansion can be used to compute inner products in the usual way.

²⁸Recall §6.5 of chapter 1, *Harmonics of the Mikrokosmos* where the quantum state Ψ of an electron trapped in a box was described in terms of harmonic functions within the boundaries of L .

It is not trivial to define infinite sums, since addition is a binary operation that we learn to extend by means of the associative law. An infinite sum can converge into a finite number, in this case the “projection” of the function f over g , but it can also diverge to infinity, depending on the values. When two functions are bounded, meaning the values of $f(x)$ remain under a certain, finite value, their inner product converges²⁹ and we can define it properly. Quantum mechanics deals precisely with bounded functions.

We generalise the definition of inner product to the case where we do not have the functions f and g in their Fourier expansion, but in their usual representation as follows:

$$\langle f|g \rangle = \sum_{j=1}^{\infty} \lambda_j \xi_j \quad \implies \quad \langle f|g \rangle = \int_D f(x)g(x)dx$$

for real-valued functions, where D is the domain of f and g , and

$$\langle f|g \rangle = \sum_{j=1}^{\infty} \lambda_j^* \xi_j \quad \implies \quad \langle f|g \rangle = \int_D f^*(x)g(x)dx$$

for the case of complex-valued functions.

The integral over the domain of f and g should not seem unfamiliar since it is reminiscent of a sum over a series of points x in the domain³⁰.

§3 Group Theory

Despite a wide variety of didactic examples of *groups* in mathematics, the truly interesting examples arise with groups of transformations. In §2.3 through §2.7 we classified linear transformations into different groups, but we used the term *group* lightly. Appendix 4: *Reminders of Algebraic Definitions* provides a formal treatment of concepts like *group*, *vector space*, etc, and justifies the usage of the term in previous sections.

Quantum mechanics is often expressed in the language of group theory; *Symmetry*, for instance, is a physical property that can be best explained by means of *invariance under a group of transformations*. The level of abstraction needed to understand group theory is not trivial, so we begin this section by building an intuition on the structure of mathematical spaces.

§3.1 Topology of \mathbb{R}^2 and \mathbb{R}^3 : Grasping an Intuition

We know from the analysis of real numbers that an interval $(a, b) \subset \mathbb{R}$ is called *open* because it does not include its boundaries, the two numbers a and b . The same interval is said to be *closed* if it does contain its two boundaries, and is

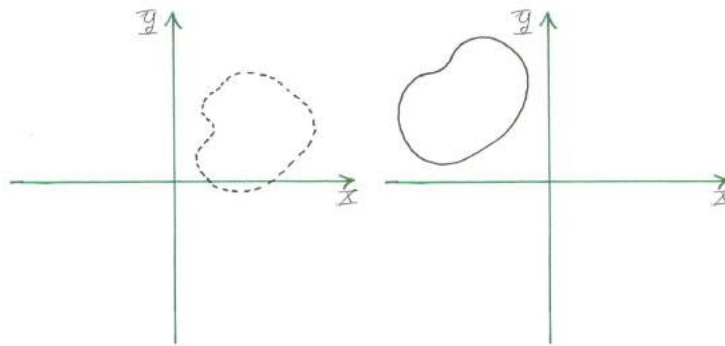
²⁹There can be pathological cases where it does not, but one can leave such cases aside in a first approach.

³⁰For an excellent introduction on Fourier analysis, see BUTKOV, E. (1968). *Mathematical physics*. Reading, Mass, Addison-Wesley Pub. Co. 18th edition.

denoted $[a, b] \subset \mathbb{R}$. An interval $[a, b)$ that is neither exactly closed nor exactly open, but somewhat both closed and open, is called a *half-open* interval.

These same notions can be extended to \mathbb{R}^2 and \mathbb{R}^3 . An open set $G \subset \mathbb{R}^2$ is *open* if every one of its points \vec{x} has a neighbourhood $\mathcal{N}_\varepsilon(\vec{x})$, centred at \vec{x} , of nearby points, lying all within some distance ε , and the entire neighbourhood is contained in the set G . Intuitively speaking, the set G is open if it does not include its boundaries, as is the case for open intervals in \mathbb{R} .

Similarly, a closed set $F \subset \mathbb{R}^2$ is such that its complement is open³¹; but, intuitively speaking, a closed set F is such that it contains its boundaries.



(a) An open set G in \mathbb{R}^2

(b) A closed set F in \mathbb{R}^2

It is useful to define these topologic properties in terms of their algebraic relations, such as *being at a distance $d < \varepsilon$ from the point \vec{x}_0* ; that way one can extend the meaning of “being open” or “being closed” to any dimension $n > 3$.

Notice that neither the open set G nor the closed set F need forcefully to be connected. Roughly speaking, a set A is connected if one can freely move between any two points $\vec{x}, \vec{y} \in A$ through a path³² without leaving A .

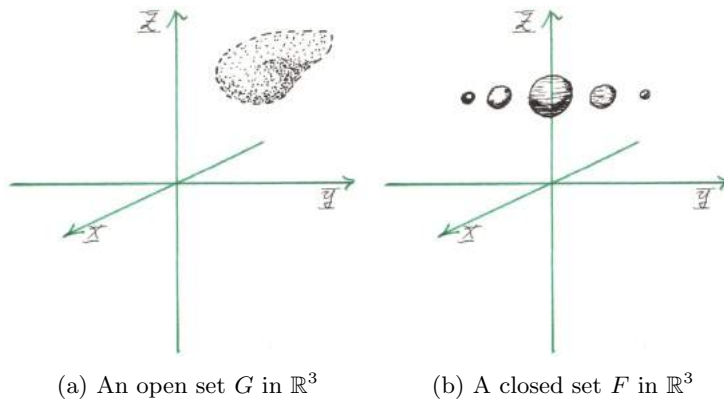
The set consisting of all open intervals $(a, b) \subset \mathbb{R}$ such that $a, b \in \mathbb{Z}$ is an example of an open, unbounded, disconnected set in \mathbb{R} . A set is bounded if it does not extend to infinity in any direction; to be specific, the set B is bounded if there exists some number $M \in \mathbb{R}$ such that for every point $\vec{x} \in B$, $\|\vec{x}\| < M$.

A set B either in \mathbb{R}^2 or \mathbb{R}^3 is said to be compact if it is both closed and bounded.

All these definitions can be extended to \mathbb{R}^n , or any other mathematical space where the notion of *distance* is already defined. The *topology* of any space is the way its open sets are related; it provides the notions of *near* and *far*, it allows us to understand the structure of mathematical spaces, and it provides

³¹It is utterly important to point out that an open (or closed) set in \mathbb{R}^2 may not necessarily be open (or closed) in \mathbb{R}^3 . I.e. the topological notions of *open* and *closed* depend tremendously on the space the set is embedded into.

³²To be precise, this is actually the definition of *path connected*, which implies connectedness, but not vice versa. For an excellent introduction to classical and functional analysis, see MARSDEN, J. E., & HOFFMAN, M. J. (1993). *Elementary Classical Analysis*. New York, W.H. Freeman.



the necessary tools to classify the sets and elements of the different spaces one happens to work with.

§3.2 Continuity: A Criterion

A function $f : A \rightarrow B$ is said to be *continuous* if it maps “close” elements $a \in A$ to “close” elements $b \in B$, where the notion of *being close* depends on each set’s topology. Notice that A and B can have different topologies, and thus quite different notions of distance. The important feature here is that, regardless of what *near* means in A , a continuous function f preserves the notion of *nearness* when mapping A into B .

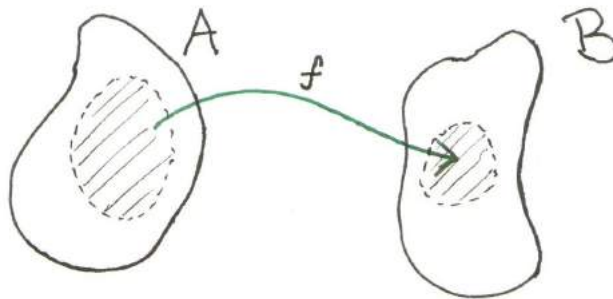


Figure 2.24: An open subset of B that comes from a subset of A under a continuous function.

This fact will turn out to be essential to classify the groups of transformations we are interested in, as well as to understand the topology of the different groups of transformations.

§3.3 The Space of Transformations

As was discussed in §2.12, the set of linear transformations between vector spaces also has a vector space structure. However, when we classified how these transformations acted upon vectors of \mathbb{R}^2 or \mathbb{R}^3 , we realised that certain groups of transformations are contained in others. For example, orthogonal transformations are part of the group of special linear transformations.

Since the determinant $\det : M_{n \times n}(\mathbb{R}) \rightarrow \mathbb{R}$ is a continuous function that maps the set of matrices into the set of real numbers, it is a useful criterion to understand how the different groups of transformations are related. For the sake of generality, we discuss the following results for \mathbb{R}^n and the groups of transformations that act upon \mathbb{R}^n . For simplicity and clarity, one can only read them as if they were about \mathbb{R}^2 or \mathbb{R}^3 .

The group $GL(n) \subset M_{n \times n}(\mathbb{R})$ is mapped under the function *determinant* onto the set of real numbers excluding 0, since for every $\hat{G} \in GL(n) \implies \det(\hat{G}) \neq 0$ as was discussed in §2.7.

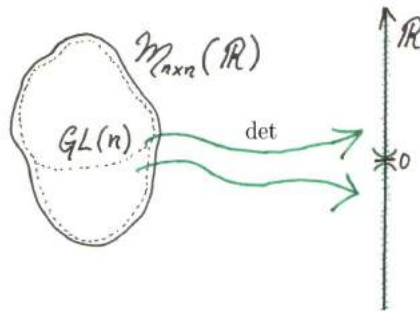


Figure 2.25: The group $GL(n)$, called *general linear group* is mapped onto the open set $\mathbb{R} - \{0\}$. Therefore, $GL(n)$ exists as an open, disconnected set within the space of linear transformations $M_{n \times n}(\mathbb{R})$.

The group $SL(n) \subset GL(n) \subset M_{n \times n}(\mathbb{R})$ is mapped under the function *determinant* into the point 1 on the line of real numbers, since for every $\hat{S} \in SL(n) \implies \det(\hat{S}) = 1$ as was discussed in §2.7.

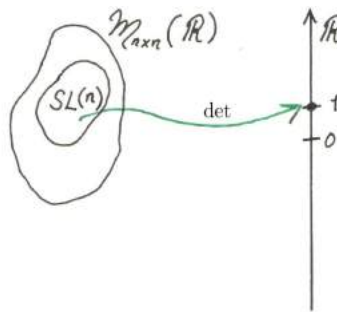
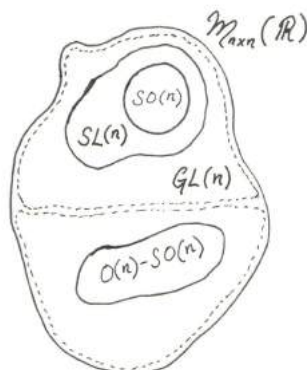


Figure 2.26: The group $SL(n)$, called *special linear group*, is mapped onto the closed set $\{1\} \in \mathbb{R}$. Therefore, $SL(n)$ exists as a closed set within the space of linear transformations $M_{n \times n}(\mathbb{R})$.

The groups $O(n), SO(n) \subset GL(n) \subset M_{n \times n}(\mathbb{R})$ of orthogonal transformations are mapped under the function *determinant* into the points 1, and -1 , depending on whether or not they preserve orientation. For every $\hat{U} \in SO(n)$ or $\hat{O} \in (O(n) - SO(n)) \implies \det(\hat{U}) = 1$ & $\det(\hat{O}) = -1$ as was discussed in §2.7.



Finally, Figure 2.28 depicts the whole structure of the space of linear transformations and the different groups it consists of. As we can see, $GL(n) \subset M_{n \times n}(\mathbb{R})$ is a disconnected, open group within the whole space; $SL(n) \subset GL(n)$ is a closed,

Figure 2.28: Topological relations within the space of linear transforma-

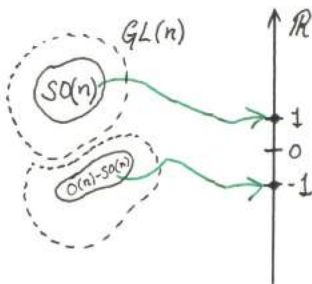


Figure 2.27: The groups $SO(n)$ and $O(n)$, called *special orthogonal group* and *orthogonal group*, are mapped onto the closed set $\{1\} \cup \{-1\} \subset \mathbb{R}$. Therefore, $O(n)$ exists as a disconnected, closed set within the space of linear transformations $M_{n \times n}(\mathbb{R})$.

connected group of transformations that preserve volumes and orientation; $SO(n) \subset O(n)$ is the group of rigid transformations that preserve orientation, volumes, and angles, it is a closed subgroup of the whole space, and “half” of the space of orthogonal transformations $O(n)$, that includes reflections; $O(n)$ is a closed, disconnected subgroup of the whole space of transformations.

§4 Probability

Probability theory is a fundamental part of quantum physics. As was deeply discussed in chapter 1, a measurement performed on any quantum system results in a value within a narrow set of possible outcomes. For a given kind of measurement there is a quantum-mechanical linear transformation that has an associated *expectation value*, a value for the state in which the system is more likely to be found. As was said before, these linear transformations are called *operators*; the term is commonly used to describe transformations that act on functional vector spaces, that map the space into itself. The derivative is an example of a linear operator; it transforms a function into another function.

Quantum theory is based upon the assumption that a given set of quantum systems *prepared* in the same state will always produce the same results for a given measurement. This probabilistic approach leaves determinism aside and assumes, axiomatically, that the actual state of a system is unknown. Most contemporary physicists suppose that there is no such thing as *the actual state of the system*, that there is nothing more to physics than its stochastic nature;

a *realist* perspective perceives a stochastic, probabilistic nature of the theory, which does not necessarily imply anything about Nature itself.

We now discuss some elementary topics of probability theory.

§4.1 Mean Value: Why is it not Enough?

We often consider average values to be representative of a sample of data; they seem to provide enough information about the overall behaviour of the system of interest. There are, however, a few problems if one works only with average values, not because they are not accurate enough, but because they contain only a limited amount of information. Life expectancy is a good example to understand the different probabilistic variables one needs in order to produce a veracious description of a system. One often reads that life expectancy during the Middle Ages was about 35, but this does not mean that people usually died at 35; actually, most people who made it to 30 could easily expect to live another 20 years. So, what is wrong with the analysis?

Suppose one wants to study life expectancy in a certain society, at a certain moment in time. One has an available set of data that relates the age of people involved in the study, and the amount of people who died at a certain age. People are divided into nine different groups consisting of the first ten years of age, then the second, and so forth. A graph like the one depicted in Figure 2.29 is called a *histogram* for this particular set of data.

We first define a *random variable* x , an association of the set of abstract possible outcomes and an actual, measurable value, usually a subset of the real numbers \mathbb{R} . In this case, our random variable is the age of death, measured in a scale from 0 to 90. Notice that the real interval $[0, 100]$ contains way too much information, for it is of no interest if someone has precisely π years of age, for example, or any other exact parameter in the set of real numbers. We approximate the age of a person within an interval Δx .

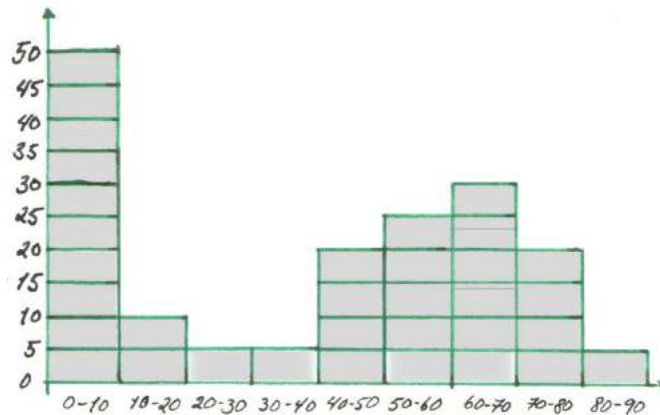


Figure 2.29: A histogram for a sample of 170 people. Fifty died between age 0 and 10, ten between 10 and 20, and so on. The Y axis presents the *frequency*, or incidence, of deaths at a certain age, which we can label $f(x)$. The X axis presents the value of the random variable x associated to age of death.

If we want to compute the *average* life expectancy in this sample, we should add up all the values of x of this particular sample, and divide it by the total amount $N = 170$. From the histogram, we do not have information on the precise age of death, but just an interval of 10 years, so we can define a *representative* parameter for each group. Without any loss of generality, we can choose the centre value; so we consider the people of the first group to have made it to the age 5, people of the second group to have made it to 15, etc. To be more succinct, we add the first value 50 times, the next one 10 times, ..., and the last value 5 times. I.e.

$$\langle x \rangle = \frac{50 \cdot (5) + 10 \cdot (15) + \dots + 5 \cdot (85)}{170} = \frac{\sum_{j=1}^9 f(x_j) \cdot x_j}{N}$$

where $\langle x \rangle$ is the average value (also called *expectation value* for experiments that

allow long-run repetitions). x_j is the representative value for each of the nine groups, $j = 1$ through $j = 9$, and $f(x_j)$ is the frequency of deaths for the value x_j . The expectation value for this sample is $\langle x \rangle = 40.29$, i.e. life expectancy for this society, at this particular moment in time, is about 40 years.

Notice how deviated this value is; the high child mortality rate affected this value drastically. So perhaps we need to define new parameters to describe this situation better. In the last equation, we divided the result of the sum by N , the total amount of people in the sample. Since N is constant, we can distribute the value $\frac{1}{N}$ in the sum, i.e.

$$\langle x \rangle = \sum_{j=1}^9 \frac{f(x_j)}{N} x_j$$

However, $\frac{f(x_j)}{N}$ is precisely the amount of occurrences for the value x_j with respect to the total N . In other words, it is a *normalised* version of this value. For example, $\frac{50}{170} = 0.2941$, then the amount of children who made it only to age 5 ± 5 represents 29.41% of the sample. So, we can define a new normalised histogram that depicts the fraction each case represents from the total 1.

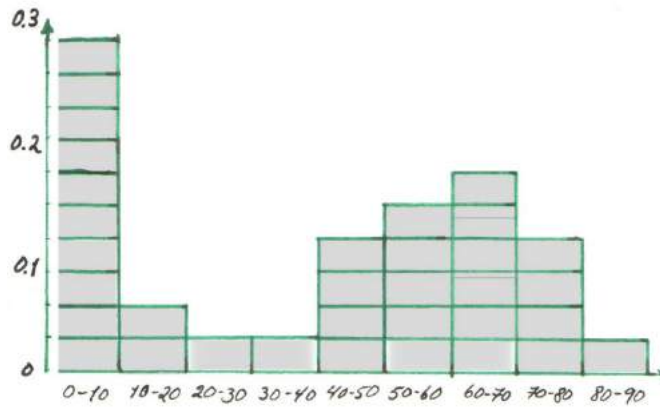


Figure 2.30: Normalised histogram, we can define a new function

$$\rho(x_j) := \frac{f(x_j)}{N} \quad \text{in that case} \quad \langle x \rangle = \sum_{j=1}^9 \rho(x_j) \cdot x_j$$

If we take one person from the sample at random, $\rho(x_j) = \frac{f(x_j)}{N}$ is the probability that such a person reached the age x_j . Adding all these probabilities individually results in

$$\sum_{j=1}^9 \rho(x_j) = \sum_{j=1}^9 \frac{f(x_j)}{N} = \frac{\sum_{j=1}^9 f(x_j)}{N} = \frac{N}{N} = 1$$

which is expected, since a random element of this sample must have died at *some* age between 0 and 90.

§4.2 Expectation Values, Variance, and Standard Deviation

A function $\rho(x)$ as the one defined above is called a probability distribution; it is a *probability mass function* for discrete variables, and *probability density function* for continuous variables. It is useful to compute other probabilistic variables, and averages of different quantities. For example, if we want to average not exactly the ages x , but some particular function $g(x)$, like the ages squared x^2 , we would have to sum the value of such function, instead of the value x_j , times the frequency for each group, the result would be

$$\langle g(x) \rangle = \sum_{j \in I} \frac{f(x_j)}{N} \cdot g(x_j) = \sum_{j \in I} \rho(x_j) \cdot g(x_j)$$

Where I is an index that depends on the amount of intervals one considers. With this in mind, we can define the new variables that will allow a better description. First, it would be useful to know how far from the average value is any given element of the sample. This is obvious if one thinks of a set of students in a school, separated into three classrooms, both of them with an average grade of 5 out of 10. One of these could be the consequence of every classmate getting a 5, while another classroom could have had half of its students with a grade very close to 10, and half with something very close to 0; the third classroom could have had half of the students with 2.5, and the other half with 7.5.

In the first case, the distribution ρ is totally concentrated at the average value 5. In the second case, there are two highly concentrated regions, namely around 0 and 10, and in the third case the distribution would show two highly concentrated regions around 2.5 and 7.5.

Whatever the case may be, we define the function $g(x_j) = x_j - \langle x \rangle$, which represents how much any element x_j differs from the average value $\langle x \rangle$. If we now average this value, we obtain

$$\begin{aligned} \langle g(x) \rangle &= \sum_{j \in I} \rho(x_j) \cdot g(x_j) = \sum_{j \in I} \rho(x_j) \cdot (x_j - \langle x \rangle) = \\ &= \sum_{j \in I} \rho(x_j) \cdot x_j - \sum_{j \in I} \rho(x_j) \cdot \langle x \rangle = \langle x \rangle - \langle x \rangle \cdot \sum_{j \in I} \rho(x_j) = \\ &= \langle x \rangle - \langle x \rangle \cdot 1 = 0 \end{aligned}$$

So the value of $g(x_j)$ was not useful, since there are as many elements that exceed the expectation value as there are elements that remain below it. We could, however, fix this problem by squaring $g(x_j)$ before computing its expectation value. We then define the *variance* $\sigma^2 := \langle (x_j - \langle x \rangle)^2 \rangle$, and we compute its expectation value as follows:

$$\begin{aligned}
\sigma^2 &= \langle (x_j - \langle x \rangle)^2 \rangle \\
&= \sum \rho(x_j) \cdot (\Delta x)^2 = \sum \rho(x_j) \cdot (x_j - \langle x \rangle)^2 \\
&= \sum \rho(x_j) \cdot x_j^2 - 2 \cdot \sum \rho(x_j) x_j \cdot \langle x \rangle + \sum \rho(x_j) \cdot \langle x \rangle^2 \\
&= \sum \rho(x_j) \cdot x_j^2 - 2 \cdot \langle x \rangle \sum \rho(x_j) \cdot x_j + \langle x \rangle^2 \cdot \sum \rho(x_j) \\
&= \sum \rho(x_j) \cdot x_j^2 - 2 \cdot \langle x \rangle \cdot \langle x \rangle + \langle x \rangle^2 \\
&= \langle x^2 \rangle - 2 \cdot \langle x \rangle^2 + \langle x \rangle^2 \\
&= \langle x^2 \rangle - \langle x \rangle^2
\end{aligned}$$

And we define the *standard deviation* σ as $\sigma := \sqrt{\sigma^2} = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}$.

These new variables, together with the expectation value, provide a much more accurate description of any statistical system.

§4.3 Distributions

This kind of discrete distributions can be adjusted to a continuous, real-valued, positive function $\rho(x)$ as is done in Figure 2.31.

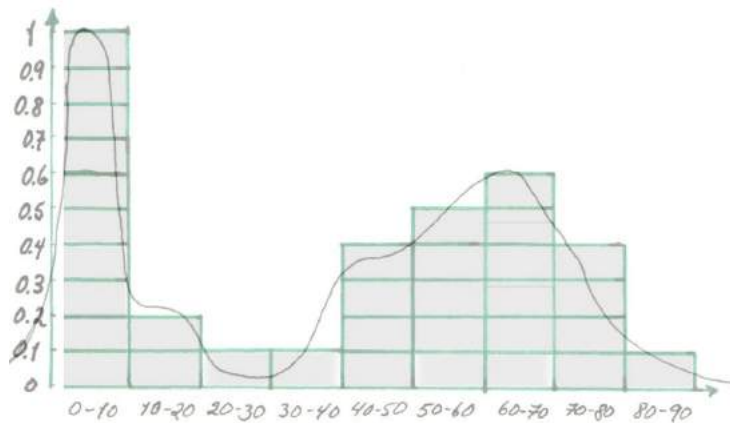


Figure 2.31: Adjusted histogram, the discrete distribution $\rho(x_j)$ is transformed into a continuous, real-valued, positive function $\rho(x)$.

And we generalise the probabilistic variables defined previously as follows:

$$\langle x \rangle = \sum_{j \in I} \rho(x_j) \cdot x_j \quad \Longrightarrow \quad \langle x \rangle = \int_{-\infty}^{\infty} \rho(x) \cdot x dx$$

and the expectation value of any function $g(x)$ as

$$\langle g(x) \rangle = \sum_{j \in I} \rho(x_j) \cdot g(x_j) \quad \Longrightarrow \quad \langle g(x) \rangle = \int_{-\infty}^{\infty} \rho(x) \cdot g(x) dx$$

where

$$\sum_{j \in I} \rho(x_j) = 1 \quad \Longrightarrow \quad \int_{-\infty}^{\infty} \rho(x) dx = 1$$

The integral over the whole domain of real numbers should seem intuitive at this point, since it represents the sum of probabilities over the whole domain³³. This is also explained in §6.2 of chapter 1, *The Size of the Wave Function: A Probability Distribution*. This last equation simply translates into “the probability that a random element has any one of the possible values is exactly 1, or 100%, equivalently.”

§5 Die Zusammenfassung

We now summarise the mathematical formalism of quantum mechanics with the aid of all the mathematical tools we have developed so far. We begin by stating the space where quantum physics is done.

§5.1 The Space of Quantum Mechanics

The wave function associated to a quantum system, as those discussed in chapter 1, is a complex valued function $\Psi : A \subset \mathbb{R}^2 \rightarrow \mathbb{C}$ that takes a pair (x, t) (in the case of one-dimensional problems) and returns a complex number $\Psi(x, t) \in \mathbb{C}$. For any given time $t = t_0$, the norm of this wave function $\Psi(x, t)$ is a probability distribution.

$$|\Psi(x)|^2 = \Psi^* \Psi = \rho(x)$$

That being the case, we impose that

$$\Psi \text{ is } \begin{cases} \cdot \text{Smooth, i.e. continuous and with continuous derivatives} \\ \cdot \text{Square integrable, i.e. its norm } |\Psi| \text{ is integrable over the whole domain.} \end{cases}$$

so that this wave function can actually represent a realistic physical system. Also,

$$\lim_{x \rightarrow \pm\infty} \Psi(x) = 0 \quad \& \quad \langle \Psi | \Psi \rangle := \int_{-\infty}^{\infty} \Psi^*(x) \Psi(x) dx$$

³³A very similar discussion is given in the book GRIFFITHS, D. J. (2005). Introduction to Quantum Mechanics. Upper Saddle River, NJ, Pearson Prentice Hall. 2nd edition.

This expression of $\|\Psi\|$ is in perfect accordance with our definition of inner product of functions. I.e. if we express $\Psi(x)$ with its Fourier expansion, then

$$\Psi(x) = \sum_{j=1}^{\infty} a_j \cdot \sin\left(\frac{2\pi}{\lambda} j \cdot x\right) + i \cdot b_j \cos\left(\frac{2\pi}{\lambda} j \cdot x\right) \quad \& \quad \|\Psi\|^2 = \sum_{j=1}^{\infty} a_j^2 + b_j^2$$

This means the space of quantum mechanics is a vector space \mathcal{H} of functions $|\Psi\rangle$ where each vector's coordinates are given by their Fourier coefficients. The notion of *metric* and *distance* between vectors is given, as usual, by the inner product $\langle\Psi|\Phi\rangle$. One must notice that, since a wave function is bounded, and the limiting value of any wave function as x approaches infinity is *zero*, all norms and their associated infinite sums are convergent. This kind of vector spaces, where any convergent series of vectors approaches elements within the vector space, i.e. not in its boundaries as defined by the space's topology³⁴, is called a *Hilbert space*.

As was deeply discussed in §6.2 of chapter 1, *The Size of the Wave Function: A Probability Distribution*, the norm squared $|\Psi(x)|^2 = \Psi^*(x)\Psi(x) = \rho(x)$ should be such that being able to find the particle *somewhere* in space is a certainty. Then,

$$\langle\Psi|\Psi\rangle = \int_{-\infty}^{\infty} \Psi^*(x)\Psi(x)dx = \int_{-\infty}^{\infty} \rho(x)dx = 1$$

Or, in terms of vectors, $\langle\Psi|\Psi\rangle = 1$, and we say the wave function Ψ is normalised.

§5.2 Expectation Values

The probabilistic definition of an expectation value, for example the average position x of a quantum particle, is given classically by

$$\langle x \rangle = \int_{-\infty}^{\infty} \rho(x) \cdot x dx$$

which is translated into the language of quantum mechanics as

$$\langle x \rangle = \int_{-\infty}^{\infty} \Psi^*(x)x\Psi(x)dx$$

The expectation value of any linear transformation \hat{T} of the vector $|\Psi(x)\rangle$ is defined analogously as

$$\langle g(x) \rangle = \int_{-\infty}^{\infty} \rho(x) \cdot g(x) dx$$

and is translated into quantum mechanical language as

$$\langle \hat{G} \rangle = \langle \Psi | \hat{G} | \Psi \rangle = \int_{-\infty}^{\infty} \Psi(x)^* \hat{G} \Psi(x) dx$$

³⁴Recall §3.1 *Topology of \mathbb{R}^2 and \mathbb{R}^3 : Grasping an Intuition*.

§5.3 The Schrödinger Equation and Eigenvalue Equations

Suppose we define a linear transformation \hat{H} that acts on the Hilbert space of wave functions $\Psi(x, t)$ as follows:

$$\hat{H} := -\frac{\hbar^2}{2m}\nabla^2 + V(x)$$

that differentiates a function twice with respect to position, and then adds a potential function $V(x)$. This linear transformation is called the *Hamiltonian* of the system; then³⁵

$$\hat{H}|\Psi\rangle = -\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2} + V(x)\cdot\Psi(x, t)$$

We then define a linear transformation associated to the *energy* of the system, acting on wave-function vectors as

$$\hat{E} := -i\hbar\frac{\partial}{\partial t}$$

that differentiates a function with respect to time, i.e

$$\hat{E}|\Psi\rangle = -i\hbar\frac{\partial\Psi}{\partial t}$$

The Schrödinger Equation can be then rewritten as

$$\hat{H}|\Psi\rangle = \hat{E}|\Psi\rangle$$

For a fixed value of energy, for example the ground state of a quantum particle E_1 ,

$$-\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2} + V(x)\cdot\Psi(x, t) = E_1\Psi(x, t)$$

We can multiply this whole equation by $\Psi^*(x, t)$ from the left and integrate over the entire domain

$$\begin{aligned} -\frac{\hbar^2}{2m}\int_{-\infty}^{\infty}\Psi^*(x, t)\frac{\partial^2\Psi}{\partial x^2}\Psi(x, t)dx + \int_{-\infty}^{\infty}\Psi^*(x, t)V(x)\cdot\Psi(x, t)dx \\ = \\ \int_{-\infty}^{\infty}\Psi^*(x, t)E_1\Psi(x, t)dx \end{aligned}$$

So

$$\langle E \rangle = \int_{-\infty}^{\infty}\Psi^*(x, t)E_1\Psi(x, t)dx = E_1\int_{-\infty}^{\infty}\Psi^*(x, t)\Psi(x, t)dx = E_1 \cdot 1$$

³⁵If this part is not clear enough, refer to *Appendix 3: Arriving at the Schrödinger Equation*.

since $\int_{-\infty}^{\infty} \Psi^*(x, t)\Psi(x, t)dx = 1$, then, the average value of this quantum particle is simply

$$\langle E \rangle = E_1$$

I.e. for a system with a fixed energy E_0 in a quantum system $|\Psi_0\rangle$ with an associated wave function $\Psi_0(x, t)$

$$\hat{H} |\Psi_0\rangle = E_0 |\Psi_0\rangle$$

In this case, E_0 is referred to as the *eigenvalue* of the linear transformation \hat{H} , and $|\Psi_0\rangle$ is referred to as the associated *eigenvector*³⁶. In general, an eigenvector is a vector that remains unchanged by the linear transformation; it is, at most, scaled by a factor λ , its eigenvalue. Eigenvectors are very useful since they provide a basis of \mathcal{H} , the vector space (Hilbert space) in which the linear transformation has an extremely simple representation, it is diagonal. In this representation, this linear transformations only “stretches” the basis vectors (in this case, the transformation’s eigenvectors). E.g.

$$\hat{T} = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}$$

In the example of a quantum particle trapped inside a box, the harmonic solutions to the Schrödinger Equation form, precisely, an orthogonal basis for the vector space of wave functions; they are eigenfunctions of the Hamiltonian operator \hat{H} . Eigenvalue equations play a major role in quantum mechanics.

³⁶or *eigenfunction* in this case

Chapter 3

An Overview of the Mathematics behind Quantum Theory



The Problem with Quantum Mechanics

§1 Interpretations

“The elegance of modern theoretical physics is largely to be found in its formal languages, not in the images with which it seeks to comprehend the world.”

Peter R. Holland ¹

A consistent quantum description of microscopic reality emerged only from the observation of collective phenomena. Its regularities are evident not on single measurements, but overall in large amounts of data, coming from ensembles of such events. Despite being an extraordinarily precise formulation, capable of predicting every single phenomenon it has been tested with, quantum theory lacks a set of notions characteristic of physical reality, such as well-defined positions and momenta of particles or their individual trajectories.

One must not, however, mistake an accurate description of physical phenomena for an explanation. Every theory must take something for granted as a starting point; at least some set of notions has to be postulated. From these apparent *ex nihilo* assumptions other physically familiar (dynamic) variables are defined, the fundamental relations are given in the language of formal mathematics, and the rest of the theory is deduced in terms of the mathematical relations one can derive from the axiomatic notions one accepts. One assumes the existence of the latter, but nothing else, since it would be unnecessary for the purposes of the theory. E.g. one does not question the existence (or even the

¹HOLLAND, P. R. (1993) *The Quantum Theory of Motion: An Account of the de Broglie-Bohm Causal Interpretation of Quantum Mechanics*. Cambridge: Cambridge University Press.

meaning) of a concept like *inertia* in Newtonian mechanics, nor is one allowed to ask what mass *is* or why it exists.

Other notions are simply left aside if their relevance to the theory or the phenomena the theory deals with is negligible. This does not mean that these notions can be neglected absolutely or discarded from all other formal descriptions; it is just clear that such a theory does not *need* these variables to predict whatever it is capable of predicting.

§1.1 Collective vs. Individual

The discrete and statistical characteristics of phenomena in the quantum realm, in contrast to the continuous and determinist characteristics of classical (macroscopic) experiments, are well described by the formalism of quantum theory. A description of the actual individuals is *not*, however, a part of its original exposition. To assume that the wave function, that which contains the information needed to predict the statistical behaviour of the system, is valid for an individual quantum particle such as an electron, is to assume that quantum mechanics is a complete description of physical reality. Thus, according to such an assumption, any notions left aside by quantum theory can and should be neglected since they belong neither to the theory nor to physical reality.

§1.2 The Problem *per se*

Thus, the main problem regarding quantum mechanics can be summarised as follows:

Given the state of a system

$$|\Psi\rangle = \sum_{j=1}^n a_j |\psi_j\rangle$$

where $\|\Psi\|^2$ is taken to be the probability density for the quantum system and each $\|a_j\|$ represents the probability contribution of the state $|\psi_j\rangle$, one can interpret this as being an intrinsic property of the quantum realm, e.g. that a quantum particle *is* actually in such a superposition of states and its existence in a certain delimited region of space is created by the act of measurement (hence the idea of a *collapse* of the wave function), with the proper probability given by the above equations; or one can, on the other hand, recognise no such attributes and interpret this in a purely statistical sense. Then, the quantum system *is* actually in a real physical state (independent of our acts of measurements) and the above equations merely represent the probabilities of the outcomes given a specific experiment.

§2 Axiomatisation: The de Broglie Hypothesis

“When I conceived the first basic ideas of wave mechanics in 1923–24, I was guided by the aim to perform a real physical synthesis, valid for all particles, of the coexistence of the wave and of the corpuscular aspects that Einstein had introduced for photons in his theory of light quanta in 1905.”

Louis de Broglie ²

As it is the case with any scientific model, quantum theory is an attempt to contain the wholeness of Nature’s intricate patterns at the atomic and subatomic scale within a set of rules, usually stated with the formality and precision of mathematical language. These rules are exactly what science look for; scientists (namely physicists) call them *physical laws*, and intend to express them as formally and unambiguously as possible. Within the scope of the mathematical language, these physical laws are called *axioms*, and they are the assumptions one makes from which everything else is to be predicted. Nothing comes *ex nihilo*, and if theory matches reality, we say the model works, and it works until Nature proves it wrong.

It is considered *elegant* to have as few laws as possible, since everything within the scope of the “theory” is to be logically inferred by means of the well-accepted schemes of mathematical logic. I.e. the set of assumptions, which can be any set of consistent starting-points, has a theoretical spectrum, a *closure of truth* consisting of every statement (theorem) that can be deduced from it. This is precisely what we formally call a *theory*. It is important to understand the formal structure of physical theories; the following examples might illustrate the point better.

In the case of Newtonian mechanics, it is the connection between the dynamic variables of the system, and the kinematic ones, what we state as an axiom (the former given in the form of a function \vec{F} , and the latter being the way momentum changes over time). The interpretation and precise meaning of the variables is related to the physical properties we have means to measure. A problem arises when the mathematical apparatus evolves faster than terminological precision. As an example, consider the concept of *mass*, which has been extensively used throughout the history of physics, and yet we have no accurate, fully-accepted definition for it.

$$\vec{F}(\vec{q}, \vec{p}, t) = \frac{d\vec{p}}{dt}$$

Newton’s Equation

In the case of thermodynamics, for example, it is from the existence of a fundamental state function and particular restraints that one derives the theory. Such a function, like the internal energy $U(S, V, \{N_i\}_{i \in I})$, contains all the thermodynamic information of the system of interest. From the way this function changes

²Louis de Broglie. DE BROGLIE, L. (1970) *The Reinterpretation of Wave Mechanics*. Vol. 1, No. 1.

one associates the following physical concepts with their formal mathematical description:

$$dU = \frac{\partial U}{\partial S} dS + \frac{\partial U}{\partial V} dV + \sum_{i \in I} \frac{\partial U}{\partial N_i} dN_i$$

where

$T := \frac{\partial U}{\partial S}$ is defined as the *Temperature*,

$P := -\frac{\partial U}{\partial V}$ is defined as the *Pressure*, and

$\mu_i := \frac{\partial U}{\partial N_i}$ is defined as the *Chemical Potential* of the substances involved.

It is relevant to notice the difference between the axioms and the derivation and definition of concepts. In the previous example, the existence of a state function containing all the physical information of a system is an axiom, but the following definitions are mere mnemotechnical devices. The theorem of equipartition of energy or Maxwell's relations are consequences of these assumptions, i.e. they can be deduced from the axioms of thermodynamics.

§2.1 Mathematics *vs.* Physics

Mathematics defines concepts, it then studies the kind of mathematical entities that might fit these definitions, regardless if they resemble the perhaps intuitive concepts that inspired such definitions. Physics deals with reality, with real objects and experimental facts; it therefore defines concepts *from* the real objects. It cannot freely explore what kind of abstract objects fit these definitions; it has to prove if the definitions actually describe the real objects. This is a fundamental difference, for physics must prioritise reality over its models. Mathematics has a certain freedom to “wander about” with concepts and definitions.

In physics, predictions are made, tested against collected data, and as long as the theory continues to serve its purpose well, it remains as a fixed paradigm. It is only when sufficiently large discrepancies arise, that the need of revision re-emerges, a new theoretical model is concocted, and the cycle starts all over again.

Quantum theory has had a remarkable resistance due to all its applications and experimental verification. Throughout its development, it has been subjected to scrutiny both by physicists and philosophers; despite its counter-intuitive nature, it continues to predict physical phenomena that is utterly unexpected, and has done so with astounding precision.

The birth of quantum mechanics can probably be traced back to de Broglie's thesis about wave properties of matter. This first *quantum conjecture* led to the development of *wave mechanics*, starting with Schrödinger's equation, and eventually the quantum theory as we know it nowadays. Even as the first inklings of debate just kindled among the physicists community, Louis de Broglie was already taking a position on the argument, stating:

“That the orthodox interpretation was not at all what I had in mind in 1923-1924 when I arrived at the idea forming the basis of wave mechanics: that the notion of coexistence of particles and waves extend to all particles. This coexistence had been discovered by Einstein in 1905 for light in his theory of photons or

light quanta. Pursuing the same course, I had been led to envisage, under the name of 'the theory of double solution,' an interpretation of wave mechanics..."

His schism with the more traditional school of thought, that of the Copenhagen Interpretation, only came formally into being after almost fifteen years of following the lead of the physicists of the time. After careful revision, he came up with the idea of a *Pilot-Wave*, a wave associated to a single particle, one that was supposed to “guide” the particle and that is only related to ψ via a scalar factor.

“The wave is, according to my notions, a physical wave of very weak amplitude whose essential role is to guide the motion of strong local concentrations of energy constituting the particles. The wave may not be arbitrarily normalised and is therefore distinct from the wave ψ , of statistical nature, utilised in quantum mechanics. I designate by v the physical wave and, in order to recover the statistical sense of the wave ψ , I define the latter by the relation $\psi = Cv$, where C is a normalisation factor. It is this essential distinction between [...] v and ψ [...] which prompted me to name my theory the 'theory of the double solution.'”

3

During the 1920's there was an avid debate on whether or not the wave function and the indeterminist nature of the theory could be a complete description of physical reality, thus revealing a true, inherent, stochastic and indeterminist nature of individual systems. De Broglie pursued to develop a more realist view by introducing both the notion of a pilot wave and a second field additional to ψ . According to his original perspective, matter waves associated to physical particles should be thought of as coexisting in physical space-time with such particles; the wave function per se was to be associated with the whole ensemble of identical particles. Such a wave function does match the probabilistic distribution of particles in space (given by $\|\psi\|^2$ naturally), but exerts also a physical influence on the particle, guiding it into specific regions according to ψ , hence the name *pilot-wave theory*.

§3 Bohmian Mechanics

By 1952, almost 25 years after the emergence of these ideas, David Bohm developed de Broglie's programme further enough explicitly to demonstrate that the assumptions of completeness⁴ in the description of individual systems (e.g. particles) is *not* a logical necessity. These results are the basis of what is often

³Louis de Broglie on a text dedicated to Alfred Landé. DE BROGLIE, L. (1971) *A New Interpretation Concerning the Coexistence of Waves and Particles*. The MIT Press, Cambridge

⁴For a deeper explanation and further understanding of the subject, see HOLLAND, P. R. (1993) *The Quantum Theory of Motion: An Account of the de Broglie-Bohm Causal Interpretation of Quantum Mechanics*. Cambridge: Cambridge University Press.

called “Bohmian Mechanics,” an interpretation of the quantum phenomena that differs from the Copenhagen Interpretation, but that predicts the exact same results in every non-relativistic experiment done so far. It is an alternative formalism that adheres to a *realistic* and deterministic perspective of quantum physics.

Bohm’s work was done independently from de Broglie’s, the former being recognised as a causal interpretation whereas the latter as a *pilot wave* theory. Both developments coincide in the realistic perspective; the theory as a whole is often called *De-Broglie-Bohm Theory of Motion*, even though they might not be equivalent *verbatim*. A brief account of its fundamental notions could be summarised with the following axioms:

- (·) A system has an associated wave that travels through space-time.
- (··) This wave ψ is a physical wave that is also a solution to the Schrödinger equation.
- (···) ψ is such that $\psi = Re^{(i/\hbar)S}$ and the particle, not just the flow, has a velocity $\vec{v} = \dot{\vec{x}} = \frac{1}{m}\nabla S$, i.e. if S is taken to be the action: the momentum corresponds to its gradient.
- (····) The probability density is still given by $\rho = \|\psi\|^2 = R^2$

Bohm tried to bring the notions of

{	• Trajectory	{	– position
	• Well defined		– momentum

back into the theory.

Once all the proper derivation is done, and Schrödinger’s equation applied, the result is a non-linear equation with an extra term interpreted as a *quantum potential*. The non-linearity of Bohm’s equations made them very unpopular at his time, and further critique was done to the fact that some stationary solutions for the Hydrogen atom imply that the velocity of an electron can, in principle, be strictly zero, which suggests a *static* model of the atom.

§4 Further Discussion (Some Final Thoughts)

§4.1 *Ignoramus et ignorabimus*

When the first set of weather-predicting equations was proposed, their non-linearity posed a major drawback. Plus, the high sensitivity to initial conditions

made any attempt to solve the problem seem futile. It was, however, this *ignorance* that prompted the sought for a new model.

Once chaos and its regularity were identified, mainly with the work of Benoit Mandelbrot and the clear recognition of nature's fractal patterns, these equations were finally seen with a different perspective. Nowadays, chaos and fractal geometry play an important role in science. It was from our ignorance that the new discoveries came into being.⁵



Figure 3.1: Paintings like Hokusai's Great Wave suggest an early, intuitive identification of fractals in nature, way before Mandelbrot's work.

The problems with quantum physics as we know it nowadays seem to be closely related to our current inability to identify and study certain properties at the atomic scale. Despite our experimental capacity, the quantum realm remains unfamiliar, even to today's physicists community. Most of our research is done separately, and different areas within atomic physics often remain uncommunicated. Mathematical languages evolve faster than our capacity to learn them, and different branches of atomic physics have such a level of specialisation, that an integral development of physics seems hardly achievable. Given the level of precision quantum theory has had in terms of prediction, its foundations are not currently being revised, and many of its notions escape our epistemological grasp. The goal of current theoretical and experimental research in quantum mechanics is usually focused in applications. Few researchers dedicate themselves to examine the foundational notions of quantum theory. It might be the task of physical research of the 21st century to pose the question if our current state of ignorance might, as has happened before, trigger an even more fruitful and solid understanding of Nature⁶.

⁵Painting:

Hokusai, Katsushika *The Great Wave* (1833). British Museum, London.

⁶"In August 1997, Max Tegmark polled 48 participants of the conference "Fundamental Problems in Quantum Theory," held at the University of Maryland, Baltimore County, about their favorite interpretation of quantum mechanics. The participants completed a questionnaire containing 16 multiple-choice questions probing opinions on quantum-foundational issues. Participants included physicists, philosophers, and mathematicians. We describe our

findings, identify commonly held views, and determine strong, medium, and weak correlations between the answers. Our study provides a unique snapshot of current views in the field of quantum foundations, as well as an analysis of the relationships between these views.” See SCHLOSSHAUER, M. et al (2013). A Snapshot of Foundational Attitudes Toward Quantum Mechanics. Stud. Hist. Phil. Mod. Phys. 44, 222-230,DOI: 10.1016/j.shpsb.2013.04.004

Chapter 4

Appendices

§1 Appendix 1: The Structure of Space and Time

There are plenty aspects involved in an holistic appreciation of physics; but certainly one of them is closely related to the $\Phi\upsilon\sigma\iota\varsigma$ (nature, essence) of space and time. The more we understand the geometry of space, i.e. its structure, the mathematical meaning of time, and the richness of the incorporated concept of *space-time*, the more sense we can find in the internal mechanics of the Universe. Some of the concepts related to space and time have, through the every day usage, grown to be so familiar that we seldom even question their true significance. Two concepts in particular, matter and waves, seem so semantically distant in the realm of human perception that one might never notice how limited to our physical scale they are. Since these definitions (and many others) fail once we leave the human scale, either by studying the *very big* or *very small*, we are forced to rethink the range of validity of our most basic notions.

§1.1 Flatland

Let us just paraphrase the idea of Edwin Abott's 1884 "Flatland"¹ to attain the exact idea of what is meant with *the structure of space*. It is the tale of a two-dimensional creature who struggles to understand the third-dimension. For the sake of pedagogical purposes, I shall take the liberty to modify the story; on the other hand, we can simply pretend we are following another character that lives in the same place as those from Abott's original story².

¹ABBOTT, E. A., & HARPER, L. M. (2010). *Flatland: a Romance of Many Dimensions*. Peterborough, Ont, Broadview Editions.

²The inspiration for this section, and a very clear and similar exposition of these subjects can be found at RUCKER, R. v. B. (1977). *Geometry, Relativity and the Fourth Dimension*. Dover Publications.

Imagine Mr. B. Square, a two dimensional figure that lives a tranquil and relatively normal life on the plane of Flatland. He, as most figures in Flatland, has a job and a house, where he lives by himself. Inhabitants of this region of Flatland know a great deal of mathematics, physics, and geometry. They recognise, specifically, all the possible directions where one could possibly travel; they know about the North, the South, the East & the West, and all possible combinations between them. They do not know, however, about the *third dimension*. We did not expect them to know about it, of course. Their bi-dimensional nature limits their understanding of what the three dimensional world might look like. It should suffice to say that, epistemologically speaking, they simply cannot possibly imagine what it *is*, or what it might be.

One evening as Mr. B. Square prepared himself to go to bed, he heard a mysterious, however clear voice coming out of nowhere. “Hello!” Said the bodiless voice. “This is Mr. A. Sphere; I am a visitor from above.” Mr. B. Square was baffled and overwhelmed by such an obstreperous intervention. How impolite it is to intrude on someone’s privacy in the middle of the night! But most importantly, *where* was this visitor?

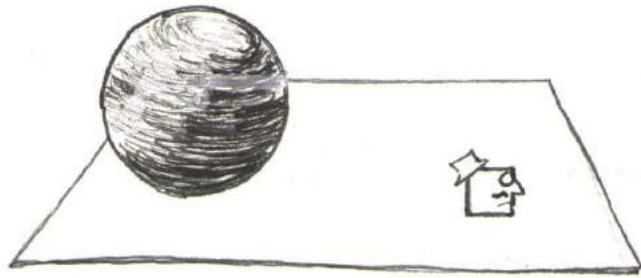


Figure 4.1: Mr. A. Sphere in Flatland

Ashamed of his behaviour, Mr. A. Sphere decided to be more specific. “I am here, good sir, just above your house.” Startled, Mr. B. Square tried to hide himself. Above? Where on Flat-Earth is this *above* thing? He obviously had never heard of such thing as “above,” and he could not see his strange visitor anywhere around. Two-dimensional beings perceive the world in 2-D, just as we perceive our own world in 3-D. They take one-dimensional pictures and paint one-dimensional portraits, just as we take and paint two-dimensional ones; they cast a one-dimensional shadow just as we cast a two-dimensional one. Their buildings, houses, and restaurants are two-dimensional boxes, just as ours are three-dimensional ones. And so Mr. A. Sphere decided to show himself by penetrating Flatland and crossing the two-dimensional plane from one side to

the other. Only then was Mr. B. Square able to “see” him. Sadly, since Mr. B. Square is a two-dimensional being, he could only see what appeared to be *slices* of Mr. A. Sphere. It was just the two-dimensional circle that results from the intersection between his world and the unexpected visitor that he was able to see, broadening and narrowing.

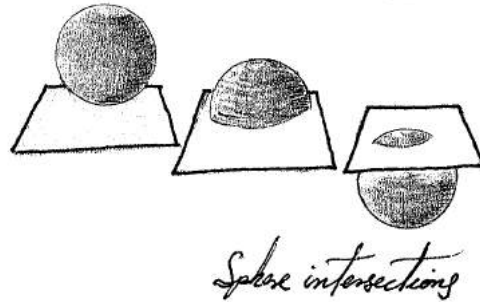


Figure 4.2: The Sphere crossing the plane

After witnessing the almost traumatising effect this intervention caused on Mr. B. Square, the three-dimensional guest decided to give a further explanation of the *third-dimension*. He told Mr. B. Square how he could see the entirety of his two-dimensional world by peacefully standing in just one spot. He told him how he could see inside all of his neighbours’ houses without even having to go inside; he did not even have to move! He even tried to explain mathematically that the third dimension is just pointing out of the plane of Flatland into a direction that is perpendicular to all possible directions Mr. B. Square could think of.

All this was just too much for poor Mr. B. Square; he never managed to appreciate fully the whole complexity of the third-dimension. He was, however, able to *deduce* what Mr. A. Sphere and his world may look like by carefully analysing and bringing together the knowledge he just recently acquired. By understanding the fact that there is much more to reality than that which he could materially experience, he was able to grasp further knowledge of the world he lived in.

The quintessential notion of this tale, the moral if you will, is the challenge that arises, this time for us, whilst seeking to imagine what is foreign to us, to our world-view. What is usually called a *hyper-sphere*, a sphere that exists in the *fourth-dimension*, challenges our own imagination, and forces us to learn from what we can deduce mathematically and leave tangible intuition behind. From its place, this fourth-dimensional sphere would be able to see us completely, without us even noticing. If we wished to point at it we would find our efforts futile, since we would have to do so in a direction that is simultaneously perpendicular to *all* directions we can think of.

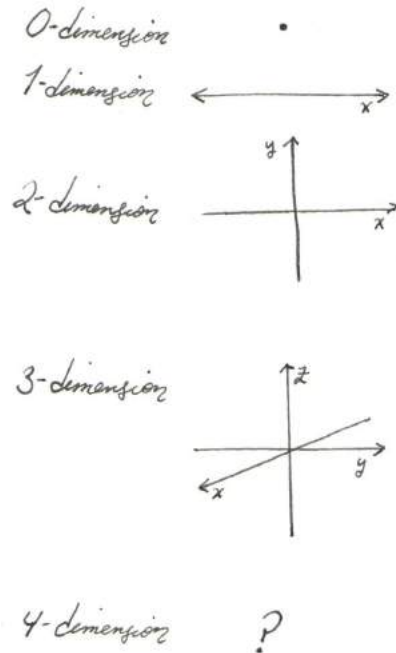


Figure 4.3: n -dimensional worlds

one dimension greater than regular space. We should be extremely careful to see that this has just a geometrical connotation; the word *dimension* has no physical meaning per se, and, therefore, no weird or “spooky” connotation.

§1.2 What *is* a Shadow?

It is easy to rush and answer this question by stating that a shadow is a manifestation of the absence of light, or perhaps even that a shadow is the darkish “image” an opaque object casts over a flat surface when exposed to a luminous source. Not often do we reflect upon the geometrical meaning of shadows, beyond our usual three-dimensional environment. Just as a bidimensional creature’s epistemic reach is bounded to a *flat* comprehension of the Universe, so is our own comprehension bounded to our three-dimensional experience.

It would have the ability to see through us, it could touch our hearts (not just figuratively) without piercing us or slicing us. It would cast a three-dimensional shadow and would be able to see all of our world just by standing still at one place. It would be able to see inside our house, and our neighbours’ without having to go inside or even move, just exactly as Mr. A. Sphere did in the *flat land*!

If it ever showed up and decided to intervene in our world the way Mr. A. Sphere did in some particular region of Flatland, i.e. by crossing from one side to the other, all we would see are the three-dimensional, sphere-like intersections of it and our world. This is how we approach the study of the fourth-dimension, a *new direction*, different to all those we know or have ever experienced. Understanding the fourth-dimension is useful (ultimately essential) to understand space, time, and the intricate relation they hold.

What we call *space-time* is always

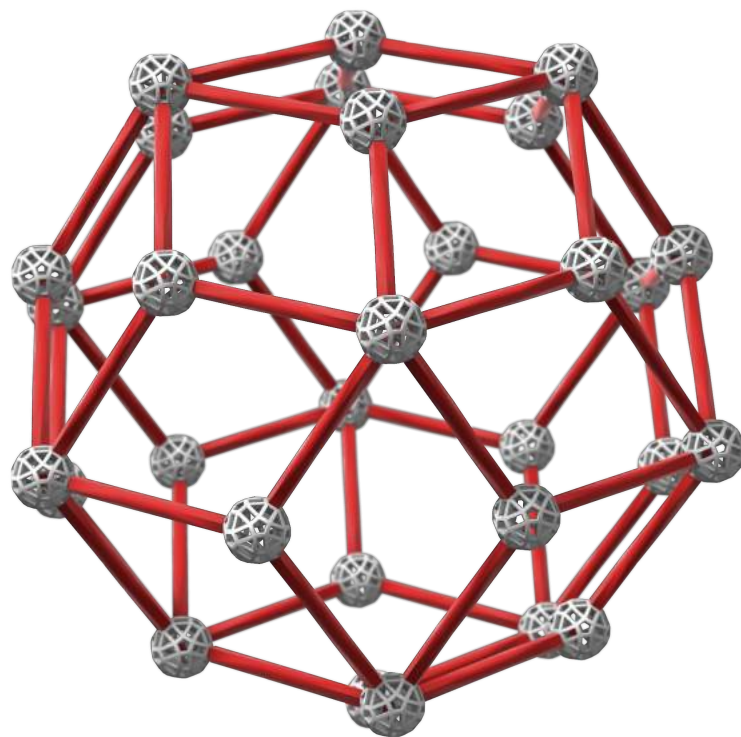
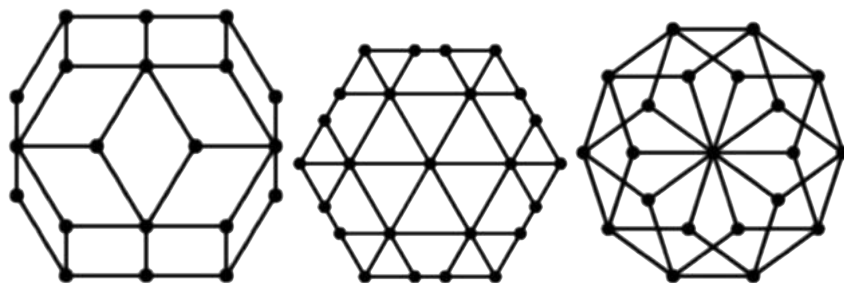


Figure 4.4: A rhombic triacontrahedron. By shining light onto it, we can see its different projections onto a plane.



(a) Flat shadow of the tri- (b) Flat shadow of the (c) Flat shadow of the tri-
 acontrahedron with two- triacontrahedron with acontrahedron with five-
 fold symmetry. three-fold symmetry. fold symmetry.

Just as the prisoners in Plato's cave allegory³, we are epistemically bound to

³More on this topic can be found on: SILVERMAN, A. "Plato's Middle Period Metaphysics and Epistemology", The Stanford Encyclopedia of Philosophy (Fall 2014 Edition), Edward N. Zalta (ed.)

think in three dimensions; four-dimensional objects can only be seen in a three-dimensional realm by means of the shadows they cast. Four or more dimensional objects can be projected into our space, and they produce a three-dimensional shadow. For us, such shadows would be indistinct from the rest of the objects around us, the same way a flat shadow would be indistinguishable from the inhabitants of the bidimensional plane.

The current physical model of our Universe comprises, precisely, a four dimensional description of our *κοσμος*, with curvature and an intrinsic hyperbolic nature. Shadows of this four dimensional world reveal only a portion, a *slice*, of the wholeness and richness of its complexity. It is easy to create false impressions from shadows, to present merely partial information of the objects they come from, and forget they only *suggest* their physical structure. In terms of Plato's allegory, it is easy to be carried away by the idea that shadows *are* the objects they represent, as a flat being might misinterpret the shadows of the rhombic triacontrahedron.

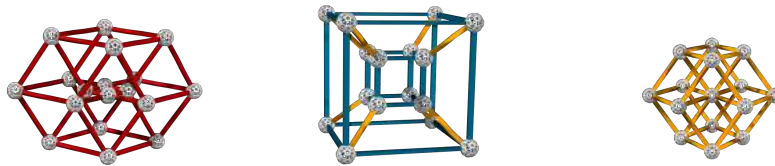


Figure 4.6: Different three-dimensional shadows of a four-dimensional cube

§1.3 *Space-Time*

In an attempt to grasp what the concept of space-time fully comprises, we start by taking a one dimensional example. Let us now consider Mr. C. Point, who lives in a one-dimensional world. He considers himself to be a point, of course, and a rather active one; we consider him nevertheless to be *zero-dimensional*, due to the way he inhabits his 1-D world. Since he is not a one-dimensional segment living in a one-dimensional world, he is not the equivalent of Mr. B. Square inhabiting Flatland (a line, not necessarily a straight one, living in Flatland would be the proper analogy). The importance of this remark shall become evident later.

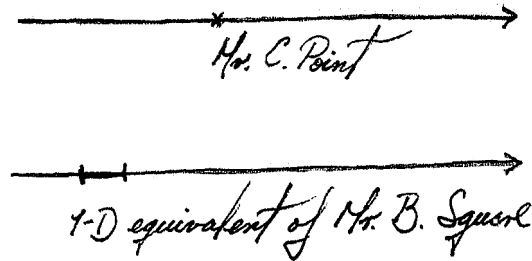


Figure 4.7: Points in the 1-D world vs 1-D people

In order to appreciate fully how Mr. C. Point moves and lives within his world we can do an interesting refashioning of the way we observe his world. He knows all the possible directions in which he can move, forwards and backwards of course. Allow me to call these directions *positive* and *negative*. His world consists, for the moment, of a straight, and infinite line. Therein lies his 1-D house and all of his possessions. By considering time, as we usually do, as a continuous parameter, as a real number⁴, we will be able to understand (and properly *see*) more of Mr. C. Point's life. We now think of the two-dimensional plane as a combination, a perpendicular intersection, of the world Mr. C. Point lives in, and time. One has to be awfully careful at this point, since in this case the plane is not to be interpreted as Flatland, but rather as an abstract depiction of the *space* we study in combination with *time* at its full extension. Space-time and space per se are two different concepts, not to be mixed up.

The following explanation of what we presently study should justify, or at least clarify, the usage of the term *space-time*. Its full profundity shall remain unexposed, but the clarity it will soon unravel is of significant importance. We will now cease to see Mr. C. Point as a point in a 1-D world (which he certainly is), and begin to see his whole life (his entire existence, his trajectory through the geography and chronology of a portion of *Lineland*, if I may call it that way) as a curve (a drawing) in the space-time diagram.

⁴Real in the set-theoretical sense, i.e. belonging to the set \mathbb{R} of real numbers

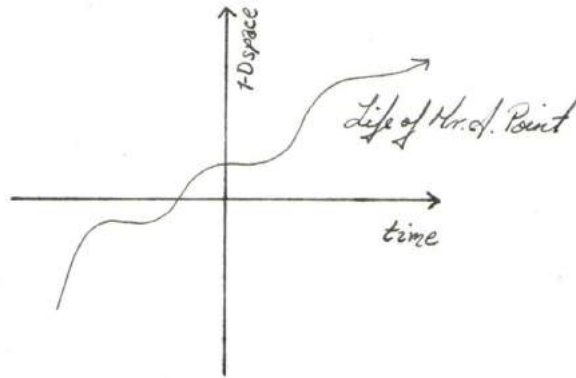


Figure 4.8: A part of Mr. C. Point's life in the Space-time diagram

No one in Lineland, or Flatland, can see the temporal dimension (direction), just as we do not see time, much less perceive it as a different direction in which we can freely travel. Inhabitants of Lineland see themselves as what they are, points. It is only we, three-dimensional beings in the attempt to study this world, that see their lives as a curve in 2-D space-time. Let us now take the abstraction one level up. Objects that, like Mr. B. Square, live in a two-dimensional world like Flatland, have a corresponding space-time as well. As one might infer from the extrapolation of our previous reasoning, this corresponding space-time is a three-dimensional one, consisting of two spatial dimensions, and one temporal one. That way, Mr. B. Square's entire life can be seen as a 3-D object, a complex, crooked, worm-like thing in the three-dimensional representation of space and time. In the diagram (Figure 4.8) we can see a portion of Mr. B. Square's life. Notice that he (and anyone around him) would only perceive himself as moving to the right. We could say that his whole world is a mere *slice* of space-time.

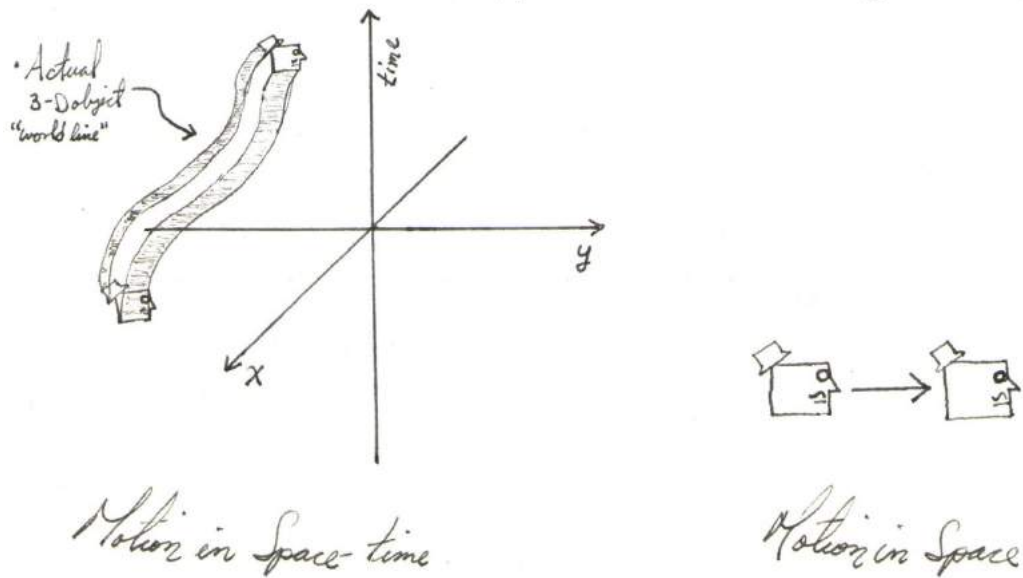


Figure 4.9: Contrast between motion in Space-Time and motion in Space

So far we know that one-dimensional space has a corresponding two-dimensional space-time; two-dimensional space has a corresponding three-dimensional space-time; from this we can clearly deduce that our own three-dimensional space has a corresponding four-dimensional space-time. It is just as impossible to draw 4-D space-time as it is to imagine it; we would have to think of a perpendicular direction to all possible directions we know, which is intrinsically impossible. We can merely describe our space-time by a 4-D arrangement (vector) with the three spatial coordinates and a temporal one. This is only a mathematical and geometrical concoction; no physical implications come intrinsically along. Let us try to go deeper into the study of the geometry of space and time, now with a slightly different orientation. Up to this point, all of our examples have been about (what we call) flat spaces, i.e. spaces with no *curvature*.

$$(t, x, y, z) \in \mathbb{R}^4$$

Mathematically speaking, an event is just a vector in \mathbb{R}^4

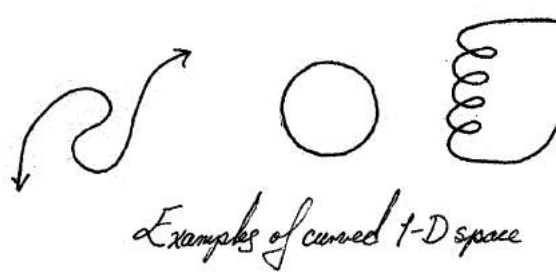
§1.4 Curvature

Let us try not to digress on this subject, but rather give a simple and short description of curved spaces, and eventually of curved *space-time*. If we recall Mr. C. Point, living in his one-dimensional world, we can see that we assumed this 1-D world (the space, not the space-time, we will forget about time for

the moment) was a straight line. This does not have to be the case, however. Notice that there is no way Mr. C. Point can feel or perceive the curvature of his space. In just the same way we (as three-dimensional beings) can see curved 1-D space, we can also see, draw, and describe curved 2-D space. Just as Mr. C. Point, however, we would not be able to perceive curvature in our own 3-D world. Curvature stretches, bends, or contracts regular “flat” space as if it were a sheet and, in the 2-D case, we can see this without any problem. But, once again, Mr. B. Square has no straight-forward way to notice the curvature in his world. The obvious question that arises is, could it be that our own three-dimensional space were curved? Is there any possible experimental way to notice this curvature? Let us first consider a few *spooky* consequences about curvature and depict some different possibilities.



Figure 4.10: Flat and curved 1-D spaces



For the sake of simplicity we leave the case of a three-dimensional space for a later examination, and focus on a two-dimensional one, like Flatland. In order to curve a two-dimensional space we need not think of it as an infinite sheet; that is an option, and in that case, space is unbounded, i.e. it has no limits. There are, nevertheless, means to curve a finite space in such way that it will remain unbounded. Let us think of different examples where 2-D space curves, but is not infinite. Certain strange properties of space appear in some of the cases. In all of the three cases illustrated below, an infinite sheet, the sphere, and the cylinder, if Mr. B. Square decided to walk East and just kept walking forever, he would never reach a limit of space. (See Figures 4.7 and 4.8)

However, if Mr. B. Square took the risk of setting course to the East in the infinite flat sheet, none of his friends and relatives would ever know of him again, since he would eventually be infinitely far away. In the sphere and the cylinder however, he would (sooner or later) come back from the other side, contradicting every reasonable intuition of flat-landers. If our Universe were the 3-D space we know, but in the form of a 3-D sphere (the equivalent of the sphere that Flatland just might be, but with three degrees of freedom of movement) we would be able to sail away (in a space ship, for instance) from the north pole on our planet Earth, into the dark emptiness of space, and eventually (many years from now, maybe) come back to Earth at the south pole. Most interestingly, as we shall study in our last two-dimensional example, is the case of the Möbius strip⁵; here, if Mr. B. Square decided to walk East, he would end up exactly where he started (just like in the sphere or the cylinder), but since the strip twists in the third-dimension, he would come back as a mirror image of himself. Remember that unlike the flat faces of three-dimensional polyhedra, objects in 2-D space have no three-dimensional orientation, i.e. creatures here have no “top” or “bottom” faces. If he was right-handed, he would now be left-handed. As a consequence, and since space is smooth and continuous, he would see everybody else as a mirror image of themselves.

What if our own space were a 3-D Möbius strip? That would mean that if two people were to travel in an interstellar cruise across the Universe, going in a particular direction (up perpendicularly to the plane of the solar system, for example, or heading towards the next galaxy but just kept going), they would end up arriving at the exact same spot where they left, here on Earth if that were the case.

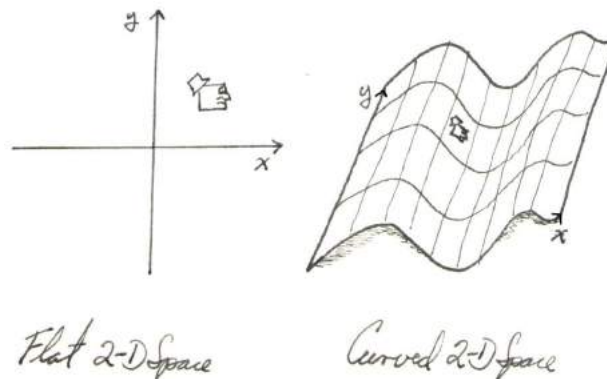


Figure 4.11: Flat and curved 2-D space

Moreover, let us suppose one of them was travelling in the seat right next to the other, to the right. Once they came back to the exact spot where they

⁵Much more and very adequate information can be found in the article “Möbius Strip” at *Wikipedia, The Free Encyclopedia* 2016

left, they would be mirrored images of themselves! They would come back left-handed (assuming they were right-handed when they left), with their hearts on the right side, and the person on the right would now be sitting on the left. That is, at least, what observers here would see; they would, however, see everyone else mirrored, and to them, the traveller originally sitting on the right seat would *still* be sitting on the right seat. This picture is not at all inconsistent; everyone sees them as mirror images of the friends who left Earth, whilst they see everyone else mirrored. This is why we say that both the Möbius strip and the Klein Bottle⁶, the three dimensional analogue of the Möbius strip, which accordingly twists in the fourth dimension, are impossible to orientate.

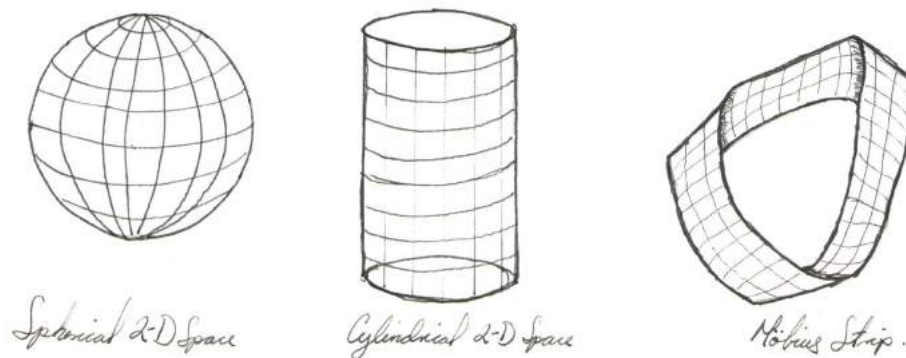


Figure 4.12: More examples of curved 2-D space. The Möbius strip is a *non-orientable* surface.

§1.5 Physics & Geometry

What would physics be like in spaces like these? Notice how straight lines *bend* in curved 2-D spaces. What could a *straight line* look like if we lived in curved 3-D spaces, which are impossible to draw or imagine? It turns out that, even though Euclid⁷ invented regular (flat) Geometry precisely to describe and study the world we live in, a weird curved space called *hyperbolic space*, proposed by Felix Klein⁸ (et al.) at the end of the 19th century as a mathematical curiosity, and further studied and applied to physics by Lorentz and Einstein⁹, turned out to be an actual closer description of reality. Two-dimensional hyperbolic spaces

⁶Much more and very adequate information can be found in the article “Klein Bottle” at *Wikipedia, The Free Encyclopedia* 2016

⁷The Euclidean description of space is the first one that is recorded. It is an attempt to formalise the knowledge of Geometry of the time; it is also, by the way, the antonomastic reference of an axiomatic system. There are several modern editions of Euclid’s (Elements) that can work as a reference.

⁸KLEIN, F. (1890). *Vorlesungen über Nicht-Euklidische Geometrie*. Göttingen.

⁹LORENTZ, H. A. (1904). “Electromagnetic Phenomena in a System Moving with any Velocity Smaller than that of Light” *Proceedings of the Royal Netherlands Academy of Arts and Science*

EINSTEIN, A. (1905). “Zur Elektrodynamik bewegter Körper” *Annalen der Physik*.

are difficult to draw; evidently, three-dimensional hyperbolic space is impossible to draw or imagine. It will suffice at the moment to say that the main characteristic of hyperbolic space is the way “straight lines” behave. The shortest path between any two points is achieved by travelling through a hyperbola.

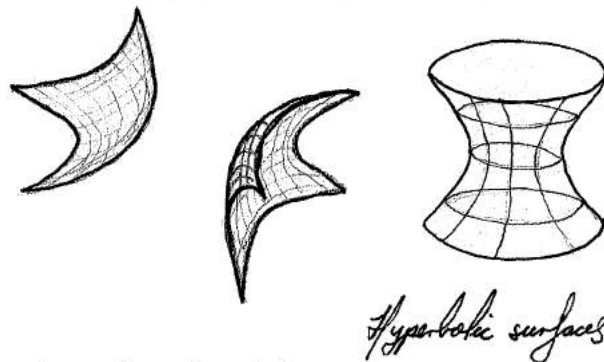


Figure 4.13: Examples of spaces with hyperbolic curvature

Einstein proposed that we, most likely, live in a *hyperbolic 4-D space-time*, which means we live in a 3-D space plus a temporal dimension; this space is curved in a hyperbolic fashion, and light travels through hyperbolae, since light has to travel in *straight lines*.

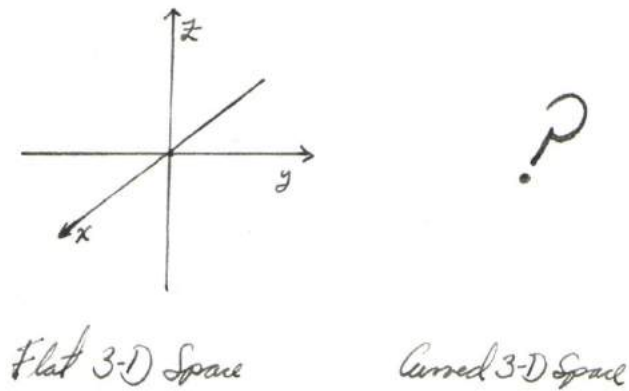


Figure 4.14: Flat and, curved 3-D space?

Another physical law he came up with, is that nothing can ever travel faster than light, not even light itself. This is remarkably counter-intuitive, and its consequences are often overseen. Not even light emitted from a travelling beam of light can travel faster than the speed of light, which stands in complete opposition to our common notions. Imagine a vehicle travelling at a considerably

high speed (relative to an observer, at least), going by just next to where the observer stands. If an object is fired from this vehicle in the same direction of motion, this object will obviously have a greater velocity than that of the vehicle, the sum of the two velocities, to be precise (with respect to the observer of course, i.e. everything measured in the same frame of references). A beam of light, on the other hand, can be projected from a rocket travelling almost at the speed of light and, both the observer standing outside, and someone travelling inside the rocket, will measure the velocity of the beam to be 2.99792458×10^8 m/s, not the sum of the two velocities in the case of the standing observer. As counter-intuitive as it may seem, everyone (travellers on that rocket, people outside, another beam of light that was in the vicinity, etc.) would record the same data for the speed of light.

We draw the limits of light and its speed with the help of a *light-cone* diagram. It is possible to make one-dimensional light-cone diagrams, as well as two-dimensional ones. Three-dimensional ones are impossible to depict, since we are dealing with space-time representations and would thus need a 4-D diagram. Hopefully, with the help of the previous explanations, the 2-D (space-time) light-cone diagram shown below will suffice. Here, light is considered to travel at a speed of 1 unit per second; this unit obviously represents 2.99792458×10^8 metres. Since nothing can possibly travel faster than light, everything that *could* happen is inside the cone. I.e. all possible events occur in the inside, the cone represents light itself, and everything outside the cone is a physical impossibility. Time goes by from bottom to top; the present lies exactly at the origin in the diagram, the past lies at the bottom, inner part of the cone, and the future lies at the upper one. Space is represented by the 2-D, XY-plane, and the shortest path between any two events (points) in hyperbolic space-time is a hyperbola. *Important fact*: if we managed to accelerate very close to the speed of light, we would move in space-time through a path very similar to a hyperbola, not an exact one. If that happened, we would arrive “faster” in the future, and we would have *travelled in time*. Another way to express this is by saying time has elapsed differently for us. This has been experimentally proved¹⁰.

¹⁰Any course on the Theory of Relativity can be useful for a better understanding of both Special and General Relativity . For a very precise and clear explanation see: COLEMAN, J. A., (1961). *Relativity for the Layman*. Penguin Books, Great Britain.

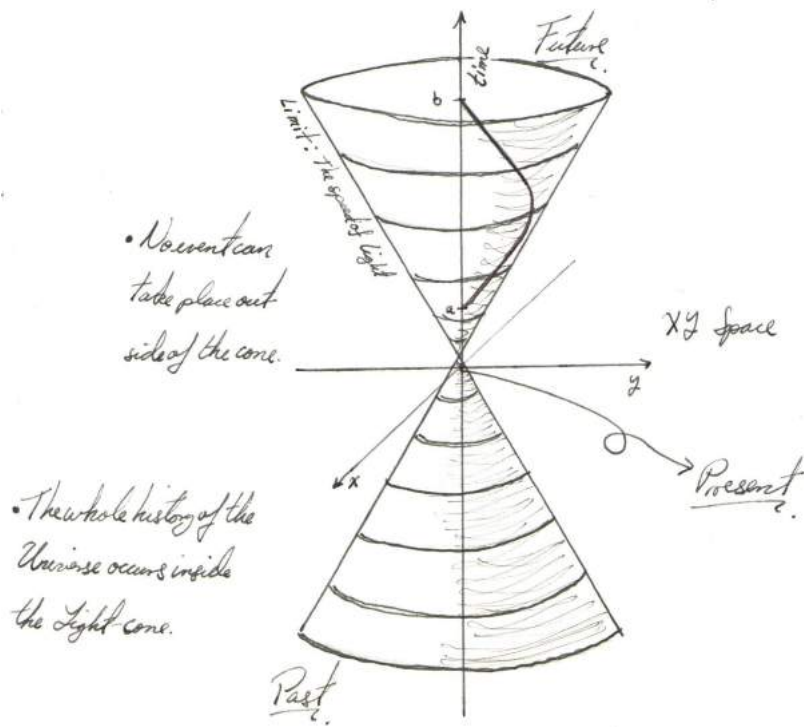


Figure 4.15: The two-dimensional Light-cone diagram

§2 Appendix 2: *Hydrodynamic Quantum Analogue, A Macroscopic Revision of Quantum Mechanics*

§2.1 What are HQA's?

A few years ago, Yves Couder and Emmanuel Fort discovered¹¹, whilst working in the laboratory, a series of phenomena that resembled properties once believed to be characteristic of quantum mechanical systems only. The experimental array they used is quite simple; by placing a droplet of a fluid (often silicone oil) over the thin layer of air that forms over the surface of a vibrating plate containing a sample of the same fluid, such a droplet is able to bounce stably and behave as a particle standing over the fluid. Although this particle-like behaviour depends strongly on the frequency of the oscillations, it is only underneath a critical acceleration under which the droplet will coalesce. So long as this threshold is exceeded, the droplet can be controlled, but only under a range of frequencies that depend on the droplet's size. This *walking* droplet has shown to behave as a quantum mechanical system, exhibiting both single and double slit diffraction, quantised orbits, quantum tunneling, and the Zeeman effect, just to name a few examples. The aim of this appendix is to present the nature of the wave-particle duality present in the quantum realm by assuming that, at least to a certain extent, these apparent manifestations of quantum effects in the macroscopic world actually portray *verbatim* how particles in the quantum realm behave.

§2.2 Fluid Mechanics

The study of wave phenomena and the study of fluid motion can hardly be separated from each other. Although an exhaustive description of the basics of hydrodynamics would be ideal, a brief description of the variables of interest and their relations will suffice. In order to understand the interrelation between fluid dynamics and the analysis of corpuscular behaviour one needs first to understand the experimental setup of this particular study. Apart from any detailed description of the experimental processes and the material used, the basic ideas underlying this kind of experiments exhibit how this connection could work. It serves as an example as well as a starting point for other research in quantum analogies.

To start with, one needs an environment, a Universe so to speak, where the droplets (which from now on shall be referred to as 'particles') can move freely as any free-particle would do under no potential constraints. The role of this Universe shall be played by a flat, rigid plate containing an oil bath. The relative size of this plate with respect to the droplet should be large enough

¹¹“Floating droplets on a vibrating bath were first described in writing by Jearl Walker in a 1978 article in Scientific American. Recently in 2005, Yves Couder and his lab were the first to systematically study the dynamics of bouncing droplets and discovered most of the quantum mechanical analogs. John Bush and his lab expanded upon Couder's work and studied the system in greater detail.” Quote from: Wikipedia contributors. “Hydrodynamic quantum analogs.” *Wikipedia, The Free Encyclopedia*, May 2018.

to consider the “particle” to be free; also, border effects can be avoided if the plate is large enough. Many different fluids would theoretically work; silicone oil, however, provides us with enough viscosity and yet good manageability so as to control the different variables easily. The oil is where the particles can move, a medium where both waves and particles can coexist.

Droplets of any fluid placed on the surface of such a fluid would normally coalesce. When the whole environment vibrates at certain specific frequencies, the time it takes for the thin layer of air between the droplet and the surface of the fluid to disappear is longer than the time it takes the fluid to bounce the droplet back into the air. This way, droplets remain as separate entities over the surface; they can behave as a standing particle, or move around the fluid depending on the relation between the bouncing frequency of the droplet and the vibrating frequency and amplitude of the plate as a whole.

A speaker connected to a function generator produces regular up-down oscillations. If properly attached to the speaker, the flat plate will vibrate accordingly and both the frequency and the amplitude of the oscillations can be regulated via the function generator. For every frequency, there is a span of amplitudes inside of which no Faraday waves appear, i.e. the fluid can oscillate steadily without any trace of standing waves at the surface.

Faraday waves¹² are nonlinear standing wave patterns that appear on the surface of any oscillating fluid constrained to a closed area. Such patterns form once a critical frequency, either from above or below, is crossed. Chladni plates¹³ operate precisely on this principle, and the symmetries formed both by the oscillating fluid and the Chladni plates are characteristic for each frequency. The fact that the frequencies at which the patterns appear constitute a discrete set of values, and that the conditions of specific frequencies and amplitudes at which the droplets are able to stand, walk or orbit around a point, also correlate to the importance of the notion of quantisation in nature.

§2.3 Setup for the Experiments

The setup for such experiments is as follows. First, a function generator is connected to an amplifier. This generator's task is to produce a single-frequency sinusoidal wave; the frequency of such a wave, as well as its amplitude, can be controlled up to a hundredth of a Hertz and a thousandth of a Volt respectively. After the signal has been amplified it travels to the speaker, where the plate containing the oil bath is attached. Once the plate starts oscillating, it is relatively easy to determine the range where Faraday waves do *not* appear. This threshold depends on both the frequency and the amplitude, and droplets are to be placed over the oil bath only within this range.

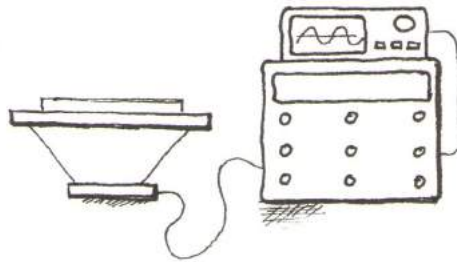


Figure 4.1: Setup for the experiment. A function generator is connected to an amplifier which is itself connected to the loudspeaker. The flat, rigid plate with the oil is fixed on top of the speaker, thus forcing it to oscillate in resonance.

¹²There is abundant literature on this very important subject of fluid mechanics. A broader and detailed explanation can be found in the following paper:
MILES, J. (1993) *On Faraday waves*, Journal of Fluid Mechanics. Cambridge University Press, 248, pp. 671–683. doi: 10.1017/S0022112093000965.

¹³For a better understanding of this phenomenon see the explanation provided by *The Science Teaching Collection* of the Smithsonian Institution. The link to it is attached below.
<http://americanhistory.si.edu/science/chladni.htm>



Figure 4.2: As the speaker reproduces the “monochromatic” wave, i.e. the single frequency coming out of the amplifier, the flat plate containing the fluid, fixed to the cone (diaphragm) of the speaker, oscillates.

§2.4 The Analogy: Connecting the Dots

The analogy works as follows, the droplet, which has been seen to stand still and move about the surface of the fluid, represents a quantum particle, such as an electron. Different “potentials,” such as single and double slits, crystal structures, corrals, etc, can be modelled using acrylic glass¹⁴, and the wave it produces whilst bouncing on the surface inevitably interacts with itself once a barrier-like object (where this wave can be reflected) is near. Depending on the amplitude of the plate’s vibrations one can make the *particle* move. If the bouncing droplet’s frequency does not match that of the plate by the smallest of phases, the droplet will be pushed around and guided by its associated wave.

This should be reminiscent of de Broglie’s *Pilot-Wave Theory*, for droplets are actual particles whose dynamics *may* obey the laws of quantum mechanics (given by the Schrödinger equation), but they also have an associated, physical wave that produces an overall wave-like behaviour.

If this actually resembles a real quantum system, we could eventually achieve a better understanding of the quantum world by developing an intuition in this kind of macroscopic systems, the results of which could permit the formulation of a new, realistic quantum theory, perhaps deprived from the interpretational problems all current models have.

¹⁴Polymethyl-methacrylate, commonly known as “Plexiglas”, a transparent thermoplastic



Figure 4.3: A particle with its associated wave standing over the fluid bath.

Trajectories of a walking droplet are well defined, however random they might be. As the droplet bounces, it creates a radial wave that propagates through the surface of the fluid and interacts with any nearby objects, e.g. another droplet, and itself. This guiding wave and its long term interaction with the particle could be the equivalent to a quantum particle interacting with itself.

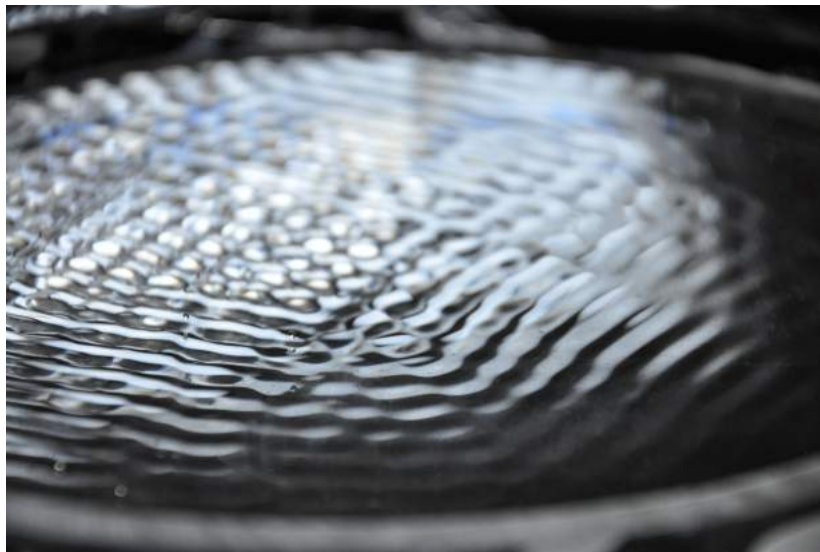


Figure 4.4: Once the Faraday threshold is exceeded, Faraday waves appear at the surface of the fluid.



Figure 4.5: Standing droplets/particles. The interference pattern is recognisable on the waves associated to the droplets on the left of the picture.



Figure 4.6: A set of droplets/particles.

§3 Appendix 3: Arriving at the Schrödinger Equation

Schrödinger's equation is a theoretical starting point in quantum mechanics, but a convincing derivation can be done with the aid of some basic physical notions. We begin by stating that the total energy of any system, be it in the classical or quantum domain, is conserved, i.e.

$$E_{Total} = E_{Kinetic} + E_{Potential}$$

where E_{Total} is a constant.

For a classical system, we defined *momentum* as $\vec{p} = m\vec{v}$, where a particle's velocity is the ratio of spatial displacement Δx over a given time interval Δt . Obviously, a fixed spatial displacement traversed over a short time interval is translated into a larger value for \vec{v} , whereas this same displacement traversed over a larger time interval translates into a smaller value for \vec{v} ¹⁵.

$$v := \lim_{\Delta t \rightarrow 0} \frac{\Delta x}{\Delta t} = \frac{dx}{dt}$$

So momentum can be seen as a physical quantity derived from the trajectory $x(t)$, both in a literal and metaphorical sense, and can thus be written as

$$\vec{p} = m \frac{d}{dt} \vec{x}(t)$$

Let us not forget that the particle's trajectory $\vec{x}(t)$ is the desired function in classical mechanics; it contains all the kinematic information of the system, and its dynamics are then given by Newton's relation $\vec{F} = m \frac{d^2}{dt^2} \vec{x}(t)$.

Finally, the system's *kinetic energy* can be re-written as

$$E_k = \frac{1}{2} m v^2 = \frac{p^2}{2m}$$

and the energy conservation statement can be written as follows:

$$\frac{p^2}{2m} + V(x) = E_{Total}$$

§3.1 Wave Function

Since we assumed any quantum system is described by a complex-valued¹⁶ wave function, we can begin with the simplest wave equation¹⁷:

$$\Psi(x, t) = e^{i(kx + \omega t)} = \cos(kx + \omega t) + i \cdot \sin(kx + \omega t)$$

¹⁵If \vec{v} is a vector, then v represents $\|\vec{v}\|$

¹⁶though sometimes real-valued

¹⁷If you are not convinced of the following equality, simply expand e^{ix} as a Taylor polynomial, i.e.

$$e^{ix} = \sum_{n=0}^{\infty} \frac{(ix)^n}{n!} = 1 + ix - \frac{x^2}{2!} - \frac{ix^3}{3!} + \dots$$

and compare it with sine's and cosine's Taylor expansion

$$\cos(x) + i \cdot \sin(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n} + i \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots + i(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots)$$

where of course

$$k = \frac{2\pi}{\lambda} \quad \& \quad \omega = \frac{2\pi}{T} = 2\pi\nu$$

and turn it into a “quantum mechanical” equation by assuming Planck’s and De Broglie’s relations

$$\lambda = \frac{h}{p} = \frac{2\pi\hbar}{p} \quad \& \quad E = \hbar\omega$$

This means that

$$k = \frac{2\pi}{\lambda} = \frac{2\pi}{\left(\frac{2\pi\hbar}{p}\right)} = \frac{p}{\hbar} \quad \& \quad \omega = \frac{E}{\hbar}$$

and the associated wave function can be re-written as

$$\Psi(x, t) = e^{i(kx + \omega t)} = e^{\frac{i}{\hbar}(px + Et)}$$

or, if it is more comfortable,

$$\Psi(x, t) = \cos\left(\frac{p}{\hbar}x + \frac{E}{\hbar}t\right) + i \cdot \sin\left(\frac{p}{\hbar}x + \frac{E}{\hbar}t\right) = \cos\left(\frac{i}{\hbar}(px + Et)\right) + i \cdot \sin\left(\frac{i}{\hbar}(px + Et)\right)$$

§3.2 Finding the Time and Spatial Derivatives

Let us find a few derivatives from this associated wave function, namely the first two spatial derivatives, and one time derivative. Recall that a function $f(x)$ that depends on just one variable has a total derivative $\frac{df}{dx}$, whereas a function $f(x, y, z)$ that depends on multiple variables has partial derivatives $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, and $\frac{\partial f}{\partial z}$, where one derives f over one variable whilst the rest of the variables are held fixed. So,

$$\frac{\partial \Psi}{\partial x} = \frac{i}{\hbar} p \cdot e^{\frac{i}{\hbar}(px + Et)}$$

$$\frac{\partial^2 \Psi}{\partial x^2} = -\frac{1}{\hbar^2} p^2 \cdot e^{\frac{i}{\hbar}(px + Et)}$$

also,

$$\frac{\partial \Psi}{\partial t} = \frac{i}{\hbar} E \cdot e^{\frac{i}{\hbar}(px+Et)}$$

This means that

$$\frac{\partial^2 \Psi}{\partial x^2} = -\frac{1}{\hbar^2} p^2 \cdot \Psi(x, t)$$

and

$$\frac{\partial \Psi}{\partial t} = \frac{i}{\hbar} E \cdot \Psi(x, t)$$

Or, equivalently

$$p^2 \Psi(x, t) = -\hbar^2 \frac{\partial^2}{\partial x^2} \Psi(x, t)$$

$$E \Psi(x, t) = \frac{\hbar}{i} \frac{\partial}{\partial t} \Psi(x, t) = -i\hbar \frac{\partial}{\partial t} \Psi(x, t)$$

Part of the great paradigm shift with quantum mechanics is the relation between physical quantities and mathematical operators. An operator is a function, a transformation, but it is commonly used to refer to functions that take functions and transform them into other functions, e.g. derivatives or integrals. Quantum mechanics is done in vector spaces, so quantum mechanical operators transform the vector space of wave functions. We associate an operator to any measurable physical quantity, like energy, momentum, and so forth. We can then take these two last equations, and define a new “quantum mechanical” *momentum operator* that is reminiscent of classical momentum, and acts upon $\Psi(x, t)$ as

$$\vec{p} = \frac{d}{dt} \vec{x}(t) \implies \hat{P} := -i\hbar \frac{\partial}{\partial x}$$

so that
$$\hat{P}^2 \Psi(x, t) = \left(i\hbar \frac{\partial}{\partial x} \right)^2 \Psi(x, t) = -\hbar^2 \frac{\partial^2}{\partial x^2} \Psi(x, t)$$

And consequently

$$E_k = \frac{1}{2}mv^2 = \frac{p^2}{2m} \implies \hat{E}_k = \frac{-\hbar^2}{2m} \frac{\partial^2}{\partial x^2} \Psi(x, t)$$

We can define the *energy operator* as

$$\hat{E} := -i\hbar \frac{\partial}{\partial t}$$

Finally, since the total energy of the system corresponds to the sum of kinetic plus potential energy, we state that

$$\frac{\hat{P}^2}{2m} + \hat{V}(x) = \hat{E}$$

where $\hat{V}(x) = V(x)$. This is an operator equation, but also

$$\left(\frac{\hat{P}^2}{2m} + \hat{V}(x) \right) \cdot \Psi(x, t) = \hat{E} \cdot \Psi(x, t)$$

where the operators are explicitly acting on $\Psi(x, t)$ from the left, and we arrive at the desired equation

$$\frac{-\hbar^2}{2m} \frac{\partial^2 \Psi}{\partial x^2} + V(x)\Psi(x, t) = -i\hbar \frac{\partial \Psi}{\partial t}$$

Or, in a more general form

$$\hat{H} |\Psi\rangle = \hat{E} |\Psi\rangle$$

where $\hat{H} = (\hat{E}_k + V(x))$ is called the *Hamiltonian operator*. This operator corresponds, in general, to the sum of kinetic and potential energy. When we analyse stationary systems, like the one described in section §6.6 of chapter 1 regarding the quantum particle trapped inside a box, we say that such systems do not evolve in time, and so we use the *time independent* version of Schrödinger's equation,

$$\frac{-\hbar^2}{2m} \frac{\partial^2 \Psi}{\partial x^2} + V(x)\Psi(x, t) = E\Psi$$

where E is the constant value for each of the different energy levels.

N.b. Other wave-like functions can be expressed as a combination of simple wave functions, like the one used in the above example, with different frequencies. In general, the associated wave function for a quantum system is a linear combination of such functions, as in the following example:

$$\begin{aligned}\Psi(\vec{x}, t) &= a_0\psi_0(\vec{x}, t) + a_1\psi_1(\vec{x}, t) + a_2\psi_2(\vec{x}, t) + \dots \\ \Psi(\vec{x}, t) &= a_0e^{i(kx+\omega_0t)} + a_1e^{i(kx+\omega_1t)} + a_2e^{i(kx+\omega_2t)} + \dots\end{aligned}$$

§4 Appendix 4: Reminders of Algebraic Definitions

The following are a few algebraic notions that are useful to understand the mathematical formalism of quantum theory. The main difference between set theory and algebra is the notion of *structure*. Sets per se have no distinction for hierarchy, order, or the way their elements relate to each other beside *set membership*, denoted as $x \in X$. Algebraic structure provides the foundations for operational mechanisms. We work with nonempty sets, and we begin with sets that have exactly one operation.

§4.1 Structures with one Binary Operation on a Set

Group

A *group* G is a set with an operation $\circ : G \times G \longrightarrow G$ such that

$$\forall g \forall h (g, h \in G \longrightarrow \circ(g, h) \in G) [Closure]$$

I.e. the group operation applied to any two elements of G is again an element of G

$$\exists e (e \in G \wedge \circ(g, e) = \circ(e, g) = g) [IdentityElement]$$

I.e. there exists a “1”

$$\forall g (g \in G \longrightarrow \exists g^{-1} (g^{-1} \in G \wedge \circ(g, g^{-1}) = e \in G)) [InverseElement]$$

I.e. for every element of G there is an inverse element (also in G) that “cancels” it out.

N.b. The group operation, represented by \circ , can be different in every case. For the integers, the operation *addition* provides them with a group structure. In that case, $\circ(g, h)$ should be interpreted as “ $m + n$ ” which is shorthand for $+(m, n)$. The identity element is the number 1, and for every $m \in \mathbb{Z}$ the inverse element is $-m$. Notice that the group operation need not be abelian (commutative).

E.g. The Rubik’s cube forms a group structure with the seven elements: Identity (doing nothing), “Quarter rotation of one side a,” “Quarter rotation of one side b,” and so on, provided we labelled each of the six sides with the letters a through f . The group operation is the composition of such quarter-rotations, i.e. performing one after the other.

In Physics, though, one usually focuses on groups of operations (actions, transformations). The set of rotations in 3-D space is a group; the group operation is again the act of successive rotations.

§4.2 Structures with two Binary Operations on a Set

Rings¹⁸

A *ring* R is a set with two operations $\oplus : R \times R \longrightarrow R$ and $\odot : R \times R \longrightarrow R$ such that

$$\forall a \forall b (a, b \in R \longrightarrow \oplus(a, b) \in R \quad \wedge \quad \odot(a, b) \in R) [Closure]$$

There exists an identity element for “multiplication,” which one can denote as “1.”¹⁹

There exists an identity element for “addition,” which one can denote as “0.”

For every element a of R there is an inverse element, with respect to the operation \odot , which one can denote $\frac{1}{a}$ (also in R) that “cancels” it out to 1.

For every element a of R there is an inverse element, with respect to the operation \oplus , which one can denote $-a$ (also in R) that “cancels” it out to 0.

“Multiplication” is associative, i.e. $(a \cdot b) \cdot c = a \cdot (b \cdot c)$

“Addition” is also associative, i.e. $(a + b) + c = a + (b + c)$

Very Important: The “addition” operation is commutative, i.e. $a + b = b + a$, but that does not imply that $a \cdot b = b \cdot a$.

Both operations combine via the *distribution law*, i.e. $a \cdot (b + c) = a \cdot b + a \cdot c$ when it is left distribution, and $(b + c) \cdot a = b \cdot a + c \cdot a$ for right distribution.

N.b. Ring-like structures are precisely characterised by the *distribution law*, $a \cdot (b + c) = a \cdot b + a \cdot c$, which tells us how the two group operations combine. A *field*, like the set of real numbers \mathbb{R} , is a special type of commutative ring structure; it is a totally ordered set with continuous parameters that is often used as a basis to construct other algebraic structures.

E.g. The set of integers, \mathbb{Z} , has a ring structure with both *multiplication* and *addition*. The set of residues, given the integer-operation “*divide by 5*,”

¹⁸“*Zahlring*,” coined by David Hilbert. In 20th century German, the word *ring* meant *association*, as in “a ring of mathematicians and philosophers.” It is still common within some contexts in modern English.

¹⁹Recall these are any arbitrary operations defined on the set; they need not be the usual addition or multiplication. Since they are so reminiscent to these familiar operations, one can simply call them by these usual names.

$\{0, 1, 2, 3, 4\}$ has a ring structure, provided we define addition the following way:

$$0 + 1 = 2$$

$$2 + 1 = 3$$

$$3 + 1 = 4$$

$$4 + 1 = 0$$

and so forth.

§4.3 Structures with two Binary Operations on two Sets

If A and F are both sets, then the two binary operations are defined as

$$\left\{ \begin{array}{l} \bullet \oplus : A \times A \longrightarrow A \\ \bullet \cdot : F \times A \longrightarrow A \end{array} \right.$$

Vector Spaces (*Linear Structures*)

A set V is a *vector space* over a field F with the operations \cdot and \oplus if

- $\cdot V$ is an abelian (commutative) group with respect to \oplus
- $\cdot F$ is a commutative ring

Which means that *addition* is closed, associative, commutative, and has both an identity element and inverse elements. *Scalar multiplication* is defined for elements in the field F and elements of V , and is a distributive operation, i.e.

There is left scalar distribution: $\lambda \cdot (v + w) = \lambda \cdot v + \lambda \cdot w$

There is right scalar distribution: $(v + w) \cdot \lambda = v \cdot \lambda + w \cdot \lambda$

N.b. Elements of vector spaces are called *vectors*, and we denote them by $|v\rangle$ or \vec{v} , depending on context. Usually, \vec{v} is used in contexts of Newtonian mechanics, where it is useful to think of vectors as arrows with length and direction; $|v\rangle$ is more useful when thinking of vectors as abstract elements of vector spaces, mainly where these are functional vector spaces.

E.g. The field \mathbb{R} per se is a vector space, but naturally $\mathbb{R}^2, \mathbb{R}^3, \mathbb{R}^4, \dots, \mathbb{R}^n$ are vector spaces. Of course,

$$\left\{ \left(\begin{array}{c} 1 \\ 0 \end{array} \right), \left(\begin{array}{c} 0 \\ 1 \end{array} \right) \right\} \text{ is a basis of } \mathbb{R}^2,$$

$\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$ is a basis of \mathbb{R}^3 ,

$\left\{ \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \dots, \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \right\}$ is a basis of \mathbb{R}^n , etc

The space of 3×3 matrices over the field of real numbers, $M_{3 \times 3}(\mathbb{R})$, is a vector space, where the 0 and 1 are given by the matrices

$$\hat{0} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \text{ and } \mathbf{1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and this space can be seen as the set of linear transformations of \mathbb{R}^3 on itself. The analogous examples for $\mathbb{R}^4, \dots, \mathbb{R}^n$ are also vector spaces of linear transformations.

§4.4 Structures with three Binary Operations on two Sets

Algebra (over a field)

A vector space V over the field \mathbb{F} is an *algebra* if it is equipped with an extra operation $\odot : V \times V \rightarrow V$ such that \odot is closed, and both right and left distribution hold, i.e.

$$u \cdot (v + w) = u \cdot v + u \cdot w$$

$$(v + w) \cdot u = v \cdot u + w \cdot u$$

N.b. This new operation (*multiplication*) is not necessarily commutative or associative.

E.g. The space of $n \times n$ matrices over the field of real (or complex) numbers, $M_{n \times n}(\mathbb{R})$ or $M_{n \times n}(\mathbb{C})$, are both algebras under *matrix multiplication*.

Of course, the set of real numbers \mathbb{R} is an algebra over itself (as a field of numbers).

§4.5 In General

If an algebra has an inner product (as the one defined in Part II), where the *length* or *norm* of its elements is defined from this inner product, and converging sequences of such elements always converge within the space itself, we call this a *Banach Algebra*.

A *Hilbert Space* is a vector space with an inner product and the corresponding topology; this vector space should also be complete, meaning that converging sequences of such elements always converge within the space itself.

A *Lie Algebra* is an example of a non-associative algebra over a field \mathbb{F} . The non-associative multiplication operation is defined with a Lie bracket $[f, g]$ such that

$$[\lambda f + \xi g, h] = \lambda[f, h] + \xi[g, h]$$

$$[f, \lambda g + \xi h] = \lambda[f, g] + \xi[f, h]$$

which means the operation is *bilinear*,

$$[f, f] = 0$$

and

$$[f, [g, h]] + [h, [f, g]] + [g, [h, f]] = 0 \quad (\text{Jacobi Identity})$$

Chapter 5

Bibliography

In order of appearance:

- [1] On Gödel's Proof (*English*) → NAGEL, E., et al (2001). *Gödel's Proof*. New York University Press
- [2] On Analytical, non-Euclidean Geometry (*Spanish*) → BRACHO, J. (2009). *Introducción analítica a las geometrías*. Fondo de Cultura Económica
- [3] Hilbert's Programme (*English*) → ZACH, R. "Hilbert's Program", *The Stanford Encyclopedia of Philosophy*, (Summer 2019 Edition), Edward N. Zalta (ed.)
- [4] Hilbert's Speech (*German*) → <https://www.math.uni-bielefeld.de/kersten/hilbert/rede.html>
- [5] Set Theory (*English*) → JECH, T., (2003). Set Theory. Springer-Verlag Berlin Heidelberg. Springer Monographs in Mathematics.
- [6] Set Theory and Infinity (*German*) → CANTOR, G. (1874), "Ueber eine Eigenschaft des Inbegriffes aller reellen algebraischen Zahlen", *Journal für die Reine und Angewandte Mathematik*, 77
- [7] Mathematical Logic (*German*) → GÖDEL, K. 1931, "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I", *Monatshefte für Mathematik und Physik*, v. 38 n. 1, pp. 173–198.
- [8] Mathematical Logic (*Spanish*) → TORRES, C. 2000, "La lógica matemática en el siglo XX", *Miscelánea Matemática SMM*, n. 31
- [9] Atomic Models (*English*) → "Atomic Theory" *Wikipedia, The Free Encyclopedia* 2016
- [10] Ancient Atomism (*English*) → BERRYMAN, S. "Democritus" *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.)

- [11] Ancient Atomism (*English*) → BERRYMAN, S. “Ancient Atomism” *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.)
- [12] Introductory Quantum Physics (*English*) → TIPLER, P. A. (1970) *Foundations of Modern Physics*. Worth Publishers, Inc. Second edition.
- [13] Introductory Quantum Physics (*English*) → BEISER, A. (2003) *Concepts of Modern Physics*. McGraw-Hill. Sixth edition.
- [14] Origins of Quantum Physics (*German*) → PLANCK, M. (1900). “Zur Theorie des Gesetzes der Energieverteilung im Normalspectrum”. *Verhandlungen der Deutschen Physikalischen Gesellschaft*.
- [15] Wave Mechanics (*French*) → DE BROGLIE, L. (1926). “Ondes et mouvements”. Solvay Conferences.
- [16] Origins of Quantum Physics (*English*) → THOMSON, G. P. (1927). “Diffraction of Cathode Rays by a Thin Film” *Nature*. 119 (3007): 890–890.
- [17] Origins of Schrödinger’s wave equation (*English*) → BLOCH, F. (1976). *Heisenberg and the Early Days of Quantum Mechanics*. *Physics Today*, 29 (12), 23-27. doi:10.1063/1.3024633
- [18] Interpretational Problem (*English*) → DE BROGLIE, L. (1971) *A New Interpretation Concerning the Coexistence of Waves and Particles*. The MIT Press, Cambridge
- [19] Quantum Physics (*English*) → BORN, M., (1969). *Atomic Physics*. Blackie, Eighth Edition.
- [20] Quantum Physics (*Spanish*) → DE LA PEÑA, L. (2003). “Introducción a la Mecánica Cuántica.” Fondo de Cultura Económica.
- [21] Electromagnetism (*English*) → GRIFFITHS, D. J., (1999). *Introduction to Electrodynamics*. Prentice Hall, Third Edition.
- [22] Uncertainty (*German*) → HEISENBERG, W. (1927). “Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik”. *Zeitschrift für Physik*.
- [23] EPR (*English*) → EINSTEIN, A., & PODOLSKY, B., & ROSEN, N. (1935). “Can Quantum-Mechanical Description of Physical Reality be Considered Complete?” *Physical Review* 47
- [24] EPR (*English*) → John Bell’s theorem. BELL, J. (1964). “On the Einstein Podolsky Rosen Paradox”. *Physics* 1
- [25] Wave Mechanics (*German*) → SCHRÖDINGER, E. (1926). “An Undulatory Theory of the Mechanics of Atoms and Molecules” *Physical Review* 28.

- [26] Particle Physics (*English*) → SRIVASTAVA, B. B., (2006), *Fundamentals of Nuclear Physics*, Rastogi Publications, India.
- [27] Linear Algebra (*English*) → FRIEDBERG, S. H., INSEL, A. J., & SPENCE, L. E. (1989). *Linear algebra*. Englewood Cliffs, N.J., Prentice Hall.
- [28] Mathematical Physics (*English*) → BUTKOV, E. (1968). *Mathematical physics*. Reading, Mass, Addison-Wesley Pub. Co. 18th edition.
- [29] Classical Analysis (*English*) → MARSDEN, J. E., & HOFFMAN, M. J. (1993). *Elementary classical analysis*. New York, W.H. Freeman.
- [30] Quantum Mechanics (*English*) → GRIFFITHS, D. J. (2005). Introduction to quantum mechanics. Upper Saddle River, NJ, Pearson Prentice Hall. 2nd edition.
- [31] Interpretational Problem (*English*) → HOLLAND, P. R. (1993) *The Quantum Theory of Motion: An Account of the de Broglie-Bohm Causal Interpretation of Quantum Mechanics*. Cambridge: Cambridge University Press.
- [32] Interpretational Problem (*English*) → DE BROGLIE, L. (1970) *The Reinterpretation of Wave Mechanics*. Vol. 1, No. 1.
- [33] De Broglie's Letter to Landau (*English*) → DE BROGLIE, L. (1971) *A New Interpretation Concerning the Coexistence of Waves and Particles*. The MIT Press, Cambridge.
- [34] Interpretations of QM Survey (*English*) → SCHLOSSHAUER, M. et al (2013). *A Snapshot of Foundational Attitudes Toward Quantum Mechanics*. Stud. Hist. Phil. Mod. Phys. 44, 222-230, DOI: 10.1016/j.shpsb.2013.04.004
- [35] Geometry and Relativity (*English*) → ABBOTT, E. A., & HARPER, L. M. (2010). *Flatland: a Romance of Many Dimensions*. Peterborough, Ont, Broadview Editions.
- [36] Geometry and Relativity (*English*) → RUCKER, R. v. B. (1977). *Geometry, Relativity and the Fourth Dimension*. Dover Publications.
- [37] Plato's Cave (*English*) → SILVERMAN, A. "Plato's Middle Period Metaphysics and Epistemology", The Stanford Encyclopedia of Philosophy (Fall 2014 Edition), Edward N. Zalta (ed.)
- [38] "Möbius Strip" → *Wikipedia, The Free Encyclopedia* 2016
- [39] "Klein Bottle" → *Wikipedia, The Free Encyclopedia* 2016
- [40] Non-Euclidean Geometry (*German*) → KLEIN, F. (1890). *Vorlesungen über Nicht-Euklidische Geometrie*. Göttingen.

- [41] Electromagnetism and Relativity (*English*) → LORENTZ, H. A. (1904). “Electromagnetic Phenomena in a System Moving with any Velocity Smaller than that of Light” *Proceedings of the Royal Netherlands Academy of Arts and Science*
- [42] Special Relativity (*German*) → EINSTEIN, A. (1905). “Zur Elektrodynamik bewegter Körper” *Annalen der Physik*.
- [43] Relativity (*English*) → COLEMAN, J. A., (1961). *Relativity for the Layman*. Penguin Books, Great Britain.
- [44] “Hydrodynamic quantum analogs.” → *Wikipedia, The Free Encyclopedia*, May 2018.
- [45] Faraday Waves (*English*) → MILES, J. (1993) *On Faraday waves*, *Journal of Fluid Mechanics*. Cambridge University Press, 248, pp. 671–683. doi: 10.1017/S0022112093000965.
- [46] Chladni Plates (*English*) → *The Science Teaching Collection*. Smithsonian Institution. <http://americanhistory.si.edu/science/chladni.htm>