SANDER BAIS

Power of the Invisible

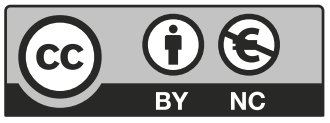
The Quantessence of Reality

Amsterdam University Press

Publication of this book has been made possible by the financial support of generous individuals as well as the following organizations and institutions:
Lorentz Fonds
Stichting Physica
Institute of Physics, University of Amsterdam
Delta ITP (NL)
Qusoft
Nikhef
Commenius Leergangen

Cover design: bij Barbara
Lay-out and illustrations: Sander Bais
This book has been typeset in Latex

ISBN 9789048562879 (cassette)
ISBN 9789048565306 (Volume I)
ISBN 9789048565313 (Volume II)
ISBN 9789048565320 (Volume III)
e-ISBN 9789048562886 (pdf)
DOI 10.5117/9789048562879
BISAC SCI057000



If quantum mechanics hasn't profoundly shocked you, you haven't understood it yet.

Niels Bohr

Whether we like it or not, modern ways are going to alter and in part destroy traditional customs and values.

Werner Heisenberg,
Physics and Philosophy:
The Revolution in Modern Science

Contents

II Quantessence: how quantum theory works

III Hierarchies: the emergence of diversity

Some like it hot! 

I assume that readers share a curiosity about quantum things, but they may have different levels of mathematical proficiency, if any. Therefore, I have put warning symbols next to the section headings. Some people like it hot & spicy and they presumably feel attracted to the sections that are marked with three hot peppers. The book is conceived such that the quite hot 2- and 3-pepper sections can be left aside without corrupting the main line of argument. *Bon appetit*!

A preface of prefaces

> We all agree that your theory is crazy. The question which divides us is whether it is crazy enough to have a chance of being correct.
>
> Niels Bohr (addressing Wolfgang Pauli)

The title *Power of the invisible* could cover a lot of possible subjects, ranging from ordinary gossip to the most elevated of spiritual teachings, as well as from the Earth's magnetic field to the invisible microcosmos. It underscores the plain fact that most things are actually invisible, unseen by the naked eye.The subtitle of this trilogy *The quantessence of reality* makes clear that in this book we limit ourselves to a world that is inaccessible to the human eye in a physical sense: A world that was only made visible hundreds of thousands of years after human history started through the development of science and technology. Humans have always been aware of the sky and the heavens, but only relatively late did they realize that there was a universe as vast, diverse and mysterious on the inside of things. The title mainly refers to the power of that hidden microcosmos, and the tremendous forces that are at work within it.

The word *quantessence* is a neologism which means 'the quintessence of quantum,' referring to phenomena that can only be explained in terms of quantum theory. A theory is a model, a symbolic representation of (a part of) the world and supposedly explains in a logically coherent way how that works. In that sense it is a visualization, an abstract reconstruction of that invisible microcosmos in terms of mathematical symbols and equations. And this is what most scientific explanations in the end tend to boil down to. And it is also this underlying network of relations and fundamental principles which govern reality that represents the power of the invisible.

The path towards such a model has been provided by an incredible interplay between science and technology, where ever more refined instruments were conceived and constructed to make discernible what was invisible before. In this way humankind has for millennia managed to push the boundaries of what is observable forward in an objective sense. And that process has fundamentally changed the nature of human existence. That is how we became aware of the tremendous *power of the invisible* and *the quantessence of reality*. The beautiful phrase 'Humans became aware, or learned about, or understood,' covers up the sobering fact that the lucky humans who are referred to unfortunately form a tiny fraction of humankind: a nerdy caste of scientist, as high priests of scientific knowledge. They are a tiny fraction in spite of the fact that everybody is invited to come and share their collective wisdom by reading books or engaging otherwise. And that turns out to be not so easy.

Scientific textbooks take pride in being as impersonal as a brick. It provides them with an aura of objectivity. Question: what do *Bethe, Baym, Bohm, Davies, Dirac, Feynman, Greiner, Griffiths, Gottfried, Kemble, Kramers, Landau, Leblond, Levy, Lipkin, Mandl, Martin, Matthews, Merzbacher, Messiah, Mott, Omnès, Pauling, Schiff, Sakurai, Shankar,Tannoudji and Weyl* have in common? Indeed, each of them has (co-)authored a textbook or two on quantum mechanics. Let me tell you how this works. If you have to teach a course on quantum theory, you can choose from more than fifty textbooks: an impressive oeuvre that bears witness to a profound love for our deepest knowledge. It doesn't stop many a teacher from adding their own little masterpiece to it. For students it is often a great relief to discover that the overlap between these books is so immense, that complete bookcases in the library effectively shrink to a tiny pile of classics. *If you've read one, you've read many.*

The personal view of the author usually becomes clear in their limited choice of subjects, and if everything is well, they should apologize for that in the Preface. That by itself

is not so exciting, in spite of being universal. Sometimes however – and that is what concerns us here – the Preface has far more to say. It appears to be the only place where the author is allowed to make their personal views known, and indeed I must admit that those are harder to embed in a treatment of, say, *angular momentum*. In the preface the author may bare their soul. It may articulate the *zeitgeist* and even deteriorate into a manifesto of principles. The innocent looking preface may actually just be a hidden persuader for personal prejudices: a *mission statement*, which might amount to little more than the scientific equivalent of what politicians call *corridor talk*. Actually, it is a place where scientist publicly tell each other the truth. Therefore this 'preface of prefaces' is a virtual quantum dialog between some of the masters which is concocted from their outspoken prefaces. This is a small quantum correction to the immaculate status of some of our quantum classics.

In 1924 the first version appeared of the standard work *Methoden der Mathematische Physik* by Courant and Hilbert (this book evolved into the monumental work in two parts that was printed in 1938). It appeared in the German university city of Göttingen at the time when the mental landslide that quantum mechanics was took place. As a matter of fact the books covered classical mathematical physics but treated the subject of differential equations and in particular eigenvalue problems in great detail, which then played a central role in solving for example the Schrödinger equation. After Courant fled Germany, long before the Second World War, the Nazi's blocked distribution of the book (as you may read in the preface to the 1953 edition). Let me share a somber quote from the original 1924 version:

> So kommt es dass viele Vertreter der Analysis das Bewusstsein der Zusammengehörigkeit ihrer Wissenschaft mit der Physik und anderen Gebieten verloren haben, während auf die andere Seite oft den Physikern das Verständnis für die Probleme and Methoden der Mathematiker, ja sogar für deren ganze Interessensphäre und Sprache abhanden gekommen ist. Ohne Zweifel liegt in dieser Tendenz eine Bedrohung für die ganze Wissenschaft überhaupt; der Strom der wissentschaftlichen Entwicklung ist in Gefahr, sich weiter und weiter zu verästeln, zu versickern und auszutrocknen.[1]

Courant therefore had no lack of drive to write a beautiful book. Another early classic (but in many ways modern) about quantum theory is *The Principles of Quantum Mechanics* by Paul Dirac (first edition in 1930). He was well-known to be a man of few words:

> Mathematics is the tool especially suited for dealing with abstract concepts of any kind and there is no limit to its power in this field. For this reason a book on the new physics, if not purely descriptive of experimental work, must be essentially mathematical.

The book then continues to present quantum theory in a form that he referred to as the 'symbolic method', a method used all over the place today:

> . . . I have chosen the symbolic method, introducing the representatives later merely as an aid to practical calculation. This has necessitated a complete break from the historical line of development, but this break is an advantage through enabling the approach to the new

[1] As a result, many practitioners of mathematical analysis have lost the awareness of their science's affiliation with physics and other fields, while on the other hand, physicists often have lost the understanding of the problems and methods of mathematicians, and indeed of their whole sphere of interest and language. There is no doubt that this trend poses a threat to the whole of science; the stream of scientific development is in danger of becoming more and more branched out, to seep away and to become dehydrated.

ideas to be made as direct as possible.

The physicists who were of the opinion that Dirac's approach was too mathematical were silenced by the quite outspoken preface of the *Mathematische Grundlagen der Quantenmechanik* by John von Neumann (1932). The opening line makes it unambiguously clear what the goals are and what the standards to be maintained throughout:[2]

> Der Gegenstand dieses Buches ist die einheitliche, und, soweit als möglich und angebracht, mathematisch einwandfreie Darstellung der neuen Quantenmechanik,...

And later on he even makes a compliment:

> Eine an Kürze und Eleganz kaum zu überbietende Darstellung der Quantenmechanik, die ebenfalls von invariantem Character ist, hat Dirac in mehreren Abhandlungen sowie in seinem kürzlich erschienenen Buche gegeben.[3]

that turns out to be a prelude to a less generous passage:

> Die erwähnte, infolge ihrer Durchsichtigkeit und Eleganz heute in einen grossen Teil der quantenmechanische Literatur übergegangene Methodik von Dirac wird den Anforderungen der mathematische Strenge in keiner Weise gerecht – auch dann nicht, wenn diese natürlicherund billigerweise auf das sonst in der theoretischen Physik übliche Mass reduziert werden.[4]

Kramers in his *Quantum Mechanics* from 1937 holds a view rather orthogonal to Von Neuman's, where he returns to the more heuristic, physically oriented approach of Bohr:

> The apparent lack of mathematical morals which is contritely pointed out repeatedly in the text is not exclusively due to the incompetence of the author. Physical morals, even (or rather especially) in their purest form, that is, unencumbered by pedagogic afterthoughts, do not live happily together with their mathematical relations in the restricted mansion of the human mind – and neither in the restricted volume of a monograph.

The famous Russian physicists Landau and Lifschitz set their own magnificent standard in their course on Theoretical Physics, which consists of more than ten volumes. These are the books from which our Russian colleagues loved to recite. If you got into a heavy-duty technical argument with them, they would exclaim: 'But don't you know this? Is well-known exercise in the chapter five, of the volume eight of the Landau Lifschitz!' Little less than the Soviet equivalent of a bible, it managed quite well to spread its profound physics wisdom. The first edition dates back to 1947. In the preface to volume three, *Quantummechanik* the authors note the following:[5]

[2] The subject of this book is the unified, and as far as possible and appropriate, mathematically rigorously correct representation of the new quantum mechanics.

[3] An account of quantum mechanics, which can hardly be surpassed in brevity and elegance, and which is also of an invariant character, has been given by Dirac in several papers as well as in his recently published book.

[4] The methodology of Dirac mentioned above, which, owing to its transparency and elegance, has today been carried over to a large part of the quantum mechanical literature, does in no way justice to the requirements of mathematical rigor, even if the standard is lowered to the more natural and reasonable one typical for theoretical physics.

[5] I apologize for quoting the German version which was for sale for a dollar or less in the former Soviet Union, at least on the rare occasions that it was not sold out. No easy reading because the formulas were set in *Fraktur* - the old German alphabet.

> Man kann nicht umhin festzustellen, dass die Darstellung in vielen Lehrbüchern der Quantenmechanik komplizierter als in den Originalarbeiten ist. Obwohl eine solche Darstellung gewöhnlich mit grösserer Allgemeinheit und Strenge begründet wird, ist jedoch bei aufmerksamer Betrachtung leicht zu erkennen, dass sowohl das eine wie die andere tatsächlich oft illusorisch sind, was sogar soweit geht, dass sich ein beträchtlicher Teil der 'strengen' Sätze als fehlerhaft erweisst. Da uns eine solche komplizierte Darstellung völlig ungerechtfertigt erscheint, haben wir uns umgekehrt um denkbar mögliche Einfachkeit bemüht und haben vielfach auf die Originalarbeiten zurückgegriffen.[6]

David Bohm also regrets in the preface to his *Quantum Theory* from 1951 the loss of qualitative, imaginable physical concepts. Bohm was well aware of the subtleties and essential role of the measurement process in quantum mechanics. And it should be said that the whole arsenal of rather puzzling, if not controversial, *Gedanken Experimente* which have in the meantime descended into the blood, sweat and tears in the lab, form a vindication of his cry to further elucidate the fundamental concepts underlying the theory:

> So strong is this contrast [between classical and quantum physics] that an appreciable number of physicists were led to the conclusion that the quantum properties of matter imply a renunciation of the possibility of their being understood in the customary imaginative sense, and that instead, there remains only a self-consistent mathematical formalism which can, in some mysterious way, predict the numerical results of actual experiments. Nevertheless, . . . , it finally became possible to express the results of the quantum theory in terms of comparitively qualitative and imaginative concepts, which are, however of a totally different nature from those appearing in the classical theory.

In this anthology we have to include the celebrated *Feynman Lectures*, as they form a most original and inspiring treatment of the theoretical basis of the physics curriculum.[7] To my knowledge it is also the first book written in first person reflecting his outspoken aversion to formality and distance. Therefore in his *Lectures* you will find regularly statements that are unmistakably Mr. Feynman like (from Part III, Chapter 1: *Quantum behavior*):

> This would mean, if it were true, that physics has given up on the problem of trying to predict exactly what will happen in a definite circumstance. Yes! Physics *has* given up.

In the preface the legendary teacher shows himself accountable for his pedagogical adventures (no need for the evaluation jungle that tends to stifle modern educational institutions):

> The question, of course, is how well this experiment succeeded. My own point of view – which, however, does not seem to be shared by most of the people who worked with the students – is pessimistic. I don't think I did well by the students. When I look at the way the

[6] One cannot help but notice that the presentation in many textbooks of quantum mechanics is more complicated than in the original works. Although such a statement is usually justified by greater generality and rigor, it is easy to see, after careful consideration, that both are often illusory, and this even goes so far as to state that a considerable part of the 'rigorous' statements prove to be faulty. As in our view such a complicated presentation appears to be completely unjustified, we have, conversely, tried to stay as simple as possible and have often resorted to the original works.

[7] The quite accessible first chapter of his book with Hibbs about *Quantum mechanics and path integrals* and his popular booklet called *QED* are also a must.

> majority of the students handled the problems on the examinations, I think that the system is a failure. . . . But then, 'The power of instruction is seldom of much efficacy except in those happy circumstances where it is almost superfluous.' (Gibbons)

There are more recent attempts to pick up the innovative approach in the presentation of quantum mechanics, for example in the book *Quantics* of Lévy-Leblond and Balibar. The term 'quantique' is apparently slang for 'quantum mechanics' used by French students. The English version 'quantics' has not seen a similar popularity among the youth educated in English, and if it is used, it is rather in in the world of data analysis and consultancy. There is a species of whizzkids called 'quants', who make money in investment banking. No quantum theory required. Yet.

> Nobody really dares to base an entire course in the spirit of these textbooks [the Feynman and Berkeley series], and often they are only used to breathe an extra bit of spirit (in some physical sense, let us say) into the traditional abstract and scholastic way of teaching. The teaching method of Feynman and Wichman is not, after all, taken seriously.

Further on in the preface we read:

> One often hears research workers expressing the desire to widen their professional culture, to deepen or rejuvenate their primary education. Such an aspiration does not come from an abstract desire to become generally cultured. Rather, it reflects the desire to increase their ability to picture, interpret and understand physics – *their* physics. To satisfy this need, these researchers all too often have at their disposal daunting and sophisticated treatises, which they find intimidating, since they have the impression that they would only find abstract answers to their concrete questions.

It was this exploration of prefaces that provided me with one of the principal motivations for writing this book. In theoretical physics and quantum theory in particular, there is always a tension between mathematical rigor and physical understanding, between formal arguments and intuition, between abstract representations and physical reality. If we look back at the development of quantum theory, we see from observational evidence that classical physics was failing us; we had to first develop a mathematical framework for the quantum world. The physical intuitions, of which the physicists were so proud, were so deeply rooted in the classical experience that they led them completely astray in the quantum world and made the development of a suitable theory very hard.

Today however, we are armed with the outcomes of a broad spectrum of real lab experiments that in the early quantum days only could be dreamt of as far-out *gedanken* experiments. There is a wide variety of quantum phenomena we have in the meantime become so 'familiar' with, that practitioners have developed a sort of *quantum intuition* – in the sense of adaption, being a healthy blend of experience and common sense. And, with that, a 'quantum heuristics' came into being – where whatever was considered esoteric speculation before, kind of turned into a bunch of 'no brainers'. This 'quantum heuristics' has at least informally gained some respectability and legitimacy. It is not quite so visible in textbooks but it is certainly predominantly present when physicists argue in front of their blackboards. I expect that this perspective will percolate through in future quantum books. One might object that this may introduce even more quantum vagueness in our quantum conversations. Apparently quantum uncertainties have made it all the way up to the heart of our ontology and epistemology, a remarkable recursion indeed.

This being said, you now know where I found the courage to produce yet another semi-popular book on quantum physics and information. You need no longer ask: 'Who ordered that?'.[8]

This book aims to demonstrate the 'Power of the invisible,' where that power refers to the 'essence' or better the 'quintessence' of quantum. This assumes that we, after more than a century of study, do know what the essence of quantum is. What we know for sure is that it is extremely powerful, in spite of being to a large extent concerned with the 'invisible.'

Talking about the essence of something requires a certain depth, not just conveying facts, but creating the appropriate reference frames and language. This *quantessential* perspective will be presented in the following Introductory chapter which also provides a roadmap to this book.

Complementary reading:

– *The Quantum Physicists*
W.H. Cropper
Oxford University. Press (1970)

[8] This is what Nobel laureate Isodor Rabi quipped in the mid-thirties, when informed about the discovery of the *muon* particle, a heavy brother of the electron that at that time seemed to have no purpose, and no reason to exist.

Further reading.
Some of the classics mentioned in this chapter:

– *Methods of Mathematical Physics*
D. Hilbert and R. Courant
Wiley-VCH; 2 Volumes (1989)

– *The Principles of Quantum Mechanics*
P.A.M. Dirac
Oxford University Press; 4th edition (1961)

– *Mathematical Foundations of Quantum Mechanics:*
J. von Neumann
Princeton Univers. Press; New edition (2018)

– *Quantum Mechanics*
H.A. Kramers
Dover Publications (1964)

– *Quantum Mechanics (Non-Relativistic Theory)*
L.D. Landau, E.M. Lifshitz
Pergamon Press; 3rd edition (1981)

– *Quantum Theory*
D. Bohm
Dover Publications Inc (1989)

– *The Feynman Lectures on Physics*
R.P. Feynman (Author), R.B. Leighton (Contributor), M. Sands (Contributor)
Pearson P T R; (3 Volume Set) 1st edition (1970)

– *Quantics: Rudiments of Quantum Physics*
J-M. Levy-Leblond F. Balibar
North Holland (1990)

Introduction

> When it comes to atoms, language can be used only as in poetry. The poet, too, is not nearly so concerned with describing facts as with creating images.
> *Niels Bohr, 'Atomic Physics and the Description of Nature' (1934)*

In this introduction we show how the book is structured and give some advice on how to read it.

Mathematics as a language of Nature. Quantum theory is known to be a difficult subject and becomes completely unfathomable if you have to rely entirely on our feeble natural language to describe it. Therefore I hope that you will not be scared away by the book's rather mathematical appearance, particularly the second volume which looks as if it is full of equations. Don't put the book aside just because of its intimidating appearance. Natural language is not the optimal means precisely because in quantum theory we enter realms of reality that are quite remote from our everyday experiences and preconceptions. Our cherished 'common sense' appeared to be of limited use and easily led us astray. Some call the quantum world mysterious or alien, while others see it as elusive or unfathomable; indeed one may easily get drowned if the message is communicated to you in words only.

Mathematics is here to rescue us; it allows us to construct smart and elegant notions that perfectly fit nature's needs and it comes with a beautifully efficient notation. The lengthy descriptions one would need in natural language to convey the essentials of quantum reality would too easily clutter the mind and lead to the utmost confusion, as I have seen happening in quite a few 'no formula' expositions of the quantum world to the layperson. So there are ample reasons to be courageous and go 'symbolic.'

The quantum leap. This art work called 'The running knot' is located in the city park of Kanazawa, Japan. (Source: Eryn Vorn, FLICKR)

Great narratives choose their own language. The heart of music is in the sound and a verbal substitute would not do. And as we all know, it takes guidance to learn how to hear what it expresses. The same is true for the visual arts. It is hard to imagine a book about Picasso without pictures. And this is what Sagredo in the *Dialogos* of Galileo confided: 'If I were again beginning my studies,

I would follow the advice of Plato and start with mathematics.' Yes, the narrative of Nature expresses itself most eloquently in mathematics. So, we take Sagredo's advice to heart and will gently introduce some of the *quantessential* mathematical concepts along the way, but always in a rather pedestrian way[9]. Math, as a language of nature, as a means for understanding, but not as a purpose on its own. To that end I have included several so-called *Math Excursions* at the end of the third volume. These excursions explain in a user-friendly way what the math in the main text is about and will tell you all you need – but maybe never wanted – to know about matters like functions, complex numbers, matrices, algebras or vectors. Checking out these excursions will help you to get more out of this book.

The best part of climbing a mountain is often the splendid view from the top. In a similar way we work our way up to some of the quantessential equations, not in praise of rigor, but in praise of clarity and beauty. I tell my students that equations love people and they better do because they owe their existence to them. Bearing that in mind, isn't it amazing that this man-made language of mathematics turns out to be the most 'natural' after all? This fascinating fact inspired the famous mathematical physicist Eugene Wigner to write an interesting essay about this paradox titled: 'The unreasonable effectiveness of mathematics in the natural sciences.' And as I intend to remain your traveling companion all along the winding road to the quantum world, I hope that you will be patient with some of the math that we will encounter along the way. Think of it as the poetry of reality: a sublime shorthand endowed with a built-in integrity. A minimal yet powerful representation of reality. There is some truth in what John von Neumann, as keynote speaker at the first national meeting of the Association for Computing Machinery in 1947, quipped: 'If

[9] As we will indeed encounter many of the fundamental equations of physics along the way, the interested reader who is not at all versed in these equations may want to look at my popular science book entitled *The Equations: icons of knowledge* (Harvard University Press, 2005).

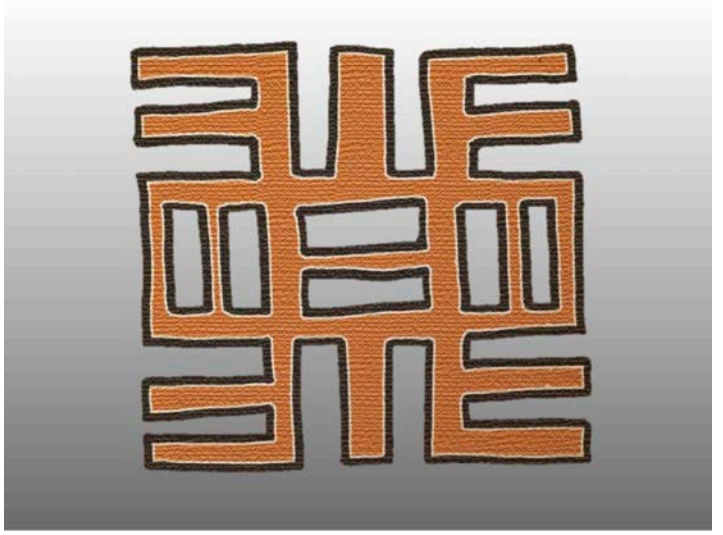


Figure 1: *Adinkra symbol.* Adinkras are symbols of the people of the Ashanti Kingdom in West Africa (Ghana) that represent concepts or wise sayings (aphorisms). This adinkra is called 'nea onnim no sua a, ohu,' which translates as 'he/she who does not know can become knowledgeable through learning.' I happen to see many interlocked copies of the letter 'E' , from Education, a striking coincidence!

S. James Gates, *Complex ideas, complex shapes* (2012)

people do not believe that mathematics is simple, it is only because they do not realize how complicated life is.'

To whom am I talking? One of the first questions a potential publisher will throw at you as potential author is about who your perceived audience is. Who is going to read (or rather, buy) this book? So many pages, so many topics, so many equations, who the hell do you think....*If you cannot kill your darlings they will kill you!* My answer is encrypted in the symbolic aphorism depicted in Figure 1 saying: *'he/she who does not know can become knowledgeable through learning'.* Keeping in mind that this holds true for basically everybody, it stands for the notion of *education permanente*, which advocates a broader spectrum of conceivable audiences for books on knowledge. There is the questionable dichotomy that books about science

have, for some reason, to belong to either the categories 'popular' or 'textbook', with basically nothing in between. From my teaching experience, I know that there are many audiences between those of laypeople and Harvard graduates. And these present us with a need for books that try to bridge the intellectual pseudo gap I just mentioned. And with the availability of internet sources like Wikipedia and Youtube there is still a clear need for in-between books that give a broad coherent account with some theoretical depth. My hope is that this book provides an example thereof. So who are the would-be members in this perceived audience: students of various backgrounds and disciplines, from motivated high school whizzkids to multidisciplinary college students, as well as their teachers. I think of students in the disciplines neighbouring physics in the natural sciences, as well as engineering, mathematics and information science. I think of journalists and of the growing group of seniors who finally have time to get to grips with some of the deep scientific subjects that over the last century through technological developments have so radically transformed the world around them. I dedicate this work to the bright young people throughout the world who share that insatiable hunger for true knowledge and I hope that it will inspire their honorable quest. Students tend to be overwhelmed by the 'how to' questions, which means that the 'why' and 'what does it mean' questions are neglected. Let me close with a quote from the early muslim polymath Al Kindi[10], who lived around 850 AD:

> We should not be ashamed of recognizing truth and assimilating it from whatever source it may reach us, even though it might come from earlier generations or foreign peoples. For him who seeks truth there is nothing of more value than truth itself. It never cheapens or abases him who searches for it, but ennobles and honors him.

[10] Al Kindi wrote more than 250 books. His Manuscript on Deciphering Cryptographic Messages, in which he laid the foundation of crypto-analysis using statistical interference and frequency analysis, is remarkable.

Nature is quantized

Quantum theory is not a theory of one particular system like the atom; it is a set of universal principles that applies to all of nature.
We present an overview of how this elaborate field is structured as a whole and thereby motivate the lay out of the book.

Quantum theory is based on a set of fundamental principles that nature appears to obey at basically all scales and therefore underlies all of physics, and more indirectly also all of chemistry and biology. The dictum is 'One Nature, One Science'. Deep down all physical theories have to behave according to the quantum rules and therefore all our theories have to be 'quantized,' somewhat like kids have to be potty-trained, and dogs have to go to obedience school to learn not to bark. The quantum postulates forced us to reinvent the whole of fundamental physics from a new conceptual basis. We have quite successfully quantized particles and mechanics, electrodynamics including optics, and liquids, solids and other condensed forms of matter. But also unified theories describing subnuclear physics have been successfully quantized and led to the celebrated *Standard Model*. And finally, not so long ago, we realized that even information should be quantized. This ongoing quantization process has led to a much deeper understanding of the fundamental structure of nature, but also to a huge number of breathtaking applications and quantum technologies that have only just started to take off. Indeed, technologies involving quantum information processing are expected to generate a highly disruptive transition with a huge socio-economic impact. Yet, this having been said, there are still many fundamental challenges, like the quantum interpretation of gravity, the oldest known force, which are required to be tackled in order to understand the origins of the universe or how black holes work.

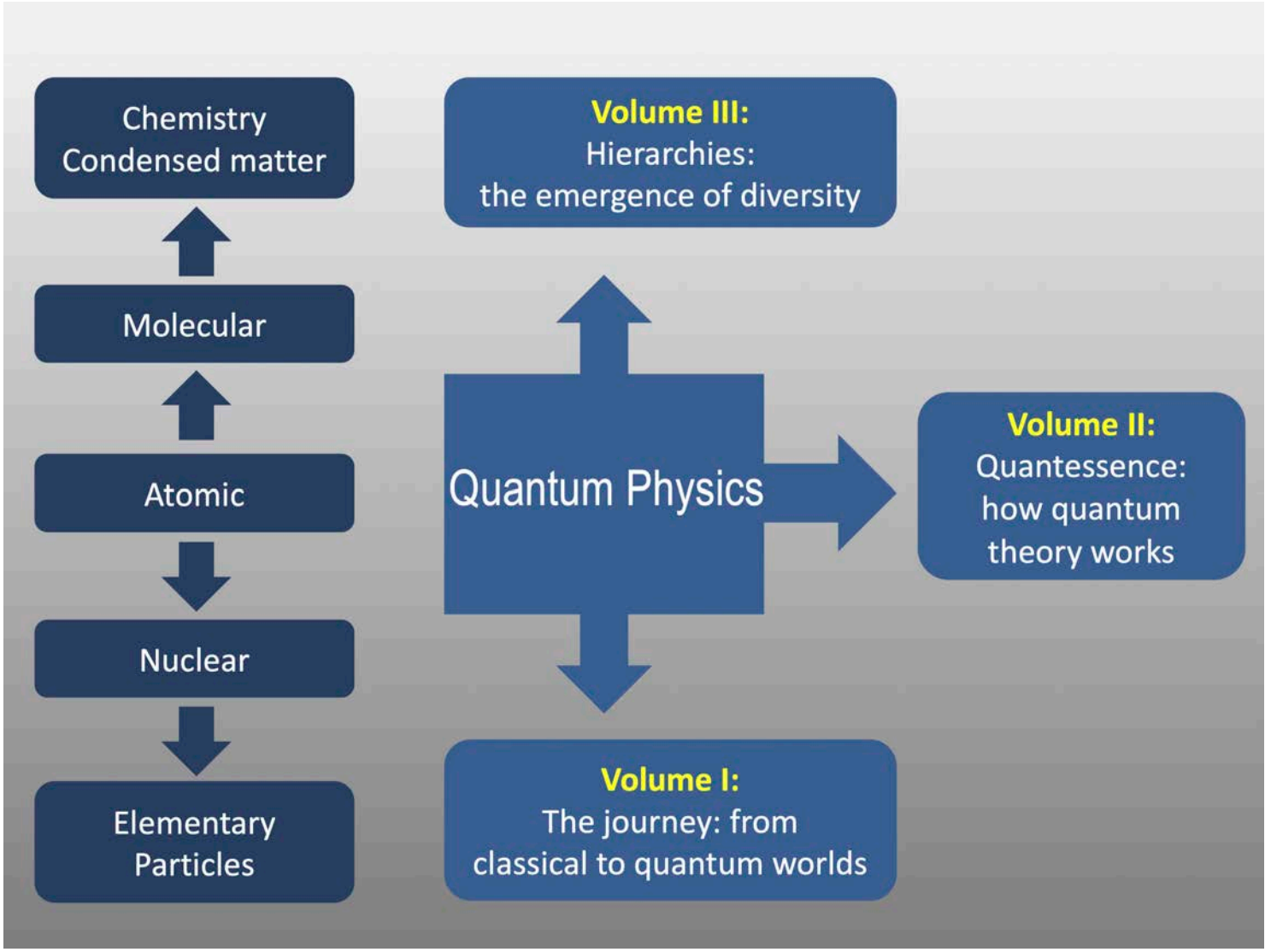


Figure 2: *Three volumes.* Quantum theory was introduced to physics at the atomic level. From there it started spreading into the other levels of physics, at both larger and smaller scales.

Three volumes. Quantum theory basically originated at the level of the atom, and by modern standards that is an intermediate length scale. From there the applications of the basic theory developed in two opposite directions, as indicated in the left column of Figure 2. On the one hand to ever smaller distances, all the way down to modern particle physics, and on the other hand to ever larger distances, moving up all the way to modern (bio)chemistry and condensed matter physics. The arrows pointing upward underscore the basic fact that quantum effects are by no means restricted to the microscopic domain. There are many research fields devoted to the study of quantum principles on macroscopic scales, which amounts to applying quantum theory to collective phenomena. In that sense every cell phone is full of quantum.

Even though I will restrict myself to the 'quantessentials', the subject is so vast that the book is divided into three parts, i.e. volumes, which – as also indicated in Figure 2 – can be characterized as follows.

The first volume of the book, called *The journey: from classical to quantum worlds*, starts with the highlights of classical physics and informatics after which it descends into the quantum world. It is the narrative guided by man's passionate quest for the most basic building blocks of nature and their interactions. We start with marbles and end up with quarks and even superstrings.

In the second volume of the book, called *Quantessence: how quantum theory works*, we delve deeper into the structure of the theory and present some of its mathematical representations. And we will talk about the conceptual issues concerning quantum states, observables and measurements that we encounter along the way. There we will be concerned extensively with mind-boggling notions like entanglement, particle interference and quantum teleportation.

In the third and final volume called *Hierarchies: the emergence of diversity*, we discuss quantum theory as it applies to the structural hierarchy of matter from the atomic level to chemistry and the quantum physics of condensed states of matter. We not only consider the hierarchy in a spatial sense but also how that hierarchy arose in a temporal sense during the early stages of cosmic evolution. It closes with a chapter on *scaling*, discussing notions such as self-similarity, scale invariance and renormalization of theories in order to understand their asymptotic behavior if one imagines the behavior of theories as models of nature, at ever smaller or larger scales. We conclude this quantum trilogy by offering a concluding chapter with a more general science-driven perspective.

Physics, mathematics and concepts

> If you look long enough, anything becomes abstract
>
> *Diane Arbus*

This section presents a meta-perspective on how to read this quantessential book. The quantum world can be traversed in many ways, all pertaining to a certain 'logic'. Taking a single path will enlighten certain aspects but may obscure others. Therefore it is better to combine different paths to get an optimal feeling for the quantum landscape. To get to the quantessence, one would have to add up the contributions of all the different paths[11].

Once a field of science (like physics) has matured sufficiently, one can learn something interesting about the structure of scientific knowledge in general. This is indicated in the layered structure of quantum knowledge in the scheme of Figure 3, in which the three columns refer to the three layers of knowledge that I like to distinguish between and which will be explained shortly.

[11] In a symbolic – if not ironic – sense, you could call this a 'path integral approach to the understanding of quantum theory.'

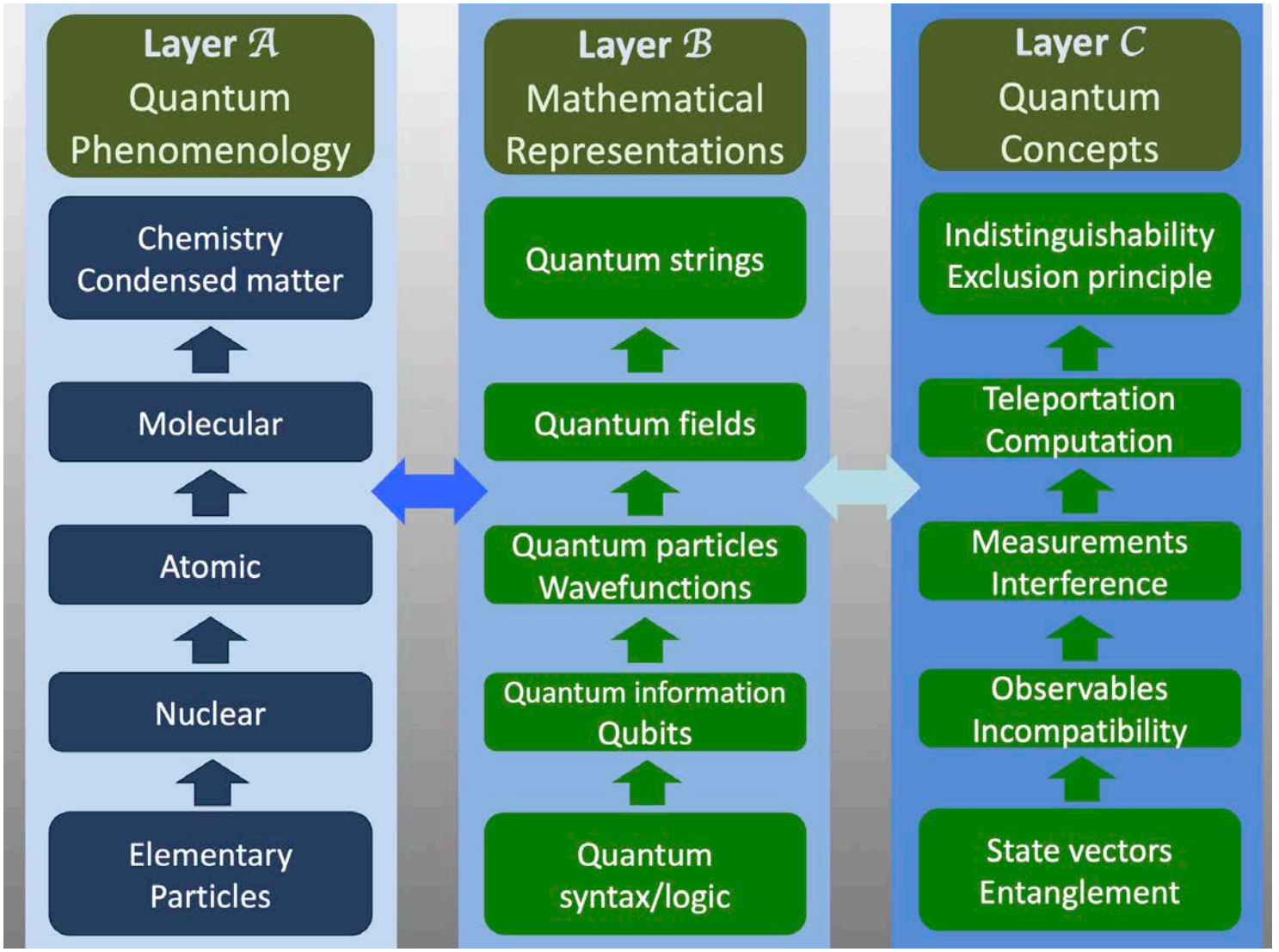


Figure 3: *Three layers.* In quantum physics one may distinguish three layers displayed here as columns. From left to right, ($\mathcal{A}$) is about the phenomenology of systems in which quantum theory manifests itself, ($\mathcal{B}$) is the layer of mathematical representations or realizations and ($\mathcal{C}$) is the layer of quantum concepts and principles. Note that the layers are coupled together as a whole, not via their individual components.

Theoretical physics is basically about constructing optimal mathematical models of reality. It usually starts by effectively describing certain regularities apparent in some observed physical phenomena. The next step – if possible – is to relate *different* phenomena through the model. This amounts to reducing the number of independent parameters in the models. Finally, one hopes that it will make predictions and suggestions as to where to look for unique signatures of new phenomena. Over time this modelling has been done in an ever more sophisticated way, exploiting existing as well as developing new mathematical and computational tools.

A first step in modelling a physical system is to just identify which degrees of freedom are relevant to the phenomena one wants to study and understand. A second crucial step is to identify what the relevant interactions between these basic degrees of freedom are. For the moment these are just words referring to basic notions, which have to find their way into some symbolic representation or mathematical framework. We may, in the end, have to extend our set of basic concepts and rules, our grammar so to speak, in order to accommodate new phenomena and new underlying principles.

In the development of quantum theory over the past century, this is exactly what happened. It turned out that we needed new mathematical realizations and ever more sophisticated representations of the material world. It is a multitude of unfolding insights intertwined with the dramatic growth of our experimental means to probe physical reality that marked the advances in theory over the last century. And finally, once the mathematical, maybe somewhat pragmatic modeling has advanced sufficiently, one should try to come to a more fundamental insight as to what these models imply for the logical structure of the underlying physical reality. Here we enter a realm with philosophical ramifications, where we move from the syntax anchored in the mathematical consistency of the model, to its semantics and interpretation. We can pose ontological questions about what is 'to be and/or not to be', as well as questions about the epistemology and about what is 'knowable'. We enter the territories of *beables* and *knowables*: in short, the realm of *meaning*.

Three layers. Quantum theory at large comprises a huge body of knowledge I like to think of as consisting of three layers as depicted in the columns of Figure 3. The $\mathcal{A}$-layer comprises the physical realizations and manifestations of quantum matter, the $\mathcal{B}$-layer is about mathematical representations and realizations, and finally the $\mathcal{C}$-layer concerns underlying concepts and principles, and their logical structure and interpretation. Indeed it is only after one has a mathematically consistent formulation of the theory that conceptual questions force themselves on us in a way that we can make sense out of them. Yet, one cannot avoid switching between the layers if one is to give a coherent account of the subject as a whole.

The first layer $\mathcal{A}$ refers to quantum phenomenology, the body of observational evidence concerning the broad spectrum of quantum phenomena that we will consider in this book. It is in fact the same as the first column in Figure 2, but note that the other columns of the two Figures refer to qualitatively entirely different things.

The second layer $\mathcal{B}$ refers to mathematical representations or models. This is already more abstract, as we ascend to the mathematical modeling of the observed phenomena. One might for example think of quantum states being elements of some vector space referred to as the Hilbert space, or of the mathematics of a qubit, or of a wave function. Or consider physical observables as represented by operators that act on that Hilbert space, like matrices or differential operators. And we may think of the dynamics of the quantum system described by famous differential equations, such as the Schrödinger, Heisenberg and Dirac equations.

And indeed, in the middle column from bottom to top we

see increasingly complex realizations of the same quantum principles, which are stated in the first step at the bottom. It is a hierarchy of degrees of freedom. We start with the discrete case of qubits and 'qubit mechanics', and move one step up to the simple continuous case of a single quantum particle. In the next step we face the problems of many particles of one single type or species, and the interactions between these species, which leads us to the theory of quantum fields. This level includes multi-particle states, and the creation and annihilation of particles; furthermore the forces are included and quantized. We finally end up with theories (and so far only theories) that attempt to combine all types of fields (or particle species) in the spectrum of a unique quantum (super)string. At this level space-time is included and quantized. So what we have indicated in the second column is the idea that states representing the physics at one given level form a small subspace of the set of states in the next step. It represents a modelling hierarchy.

We have mapped the system onto a mathematical model that allows us to make calculations and predictions, but models also pose new challenges for finding out what the essential concepts are that underlie all those quantum phenomena. We like to understand what the generic features are that set the quantum world apart from what we were used to in classical physics. That is what the next layer is about.

The third layer $\mathcal{C}$ is concerned with the conceptual implications of the mathematical framework, where we are required to interpret the basic mathematical entities back into physical terms. You may compare this to coming home from an exciting journey to some unknown country, and being forced to describe to your colleagues what the exquisite, extremely exotic food tasted like. You may think of mathematical models that manage to successfully describe and predict measurement outcomes, but at the same time force us to reinterpret what the very nature of physical reality is. There is the saying cherished by many theorists that 'equations speak for themselves', but that is often not the case. For example, you may know that the mathematics of special relativity is surprisingly simple, but its physical ramifications are certainly *not*; it forced us to fundamentally redefine our concepts of space and time. Something similar happened in the realm of quantum theory with respect to the true nature of what we, for convenience, call 'matter', or 'radiation', or 'energy', or 'information'. Here we encounter the necessary consequences of the Hilbert space formalism, such as the existence of quantum entanglement and quantum interference. And we have to cope with non-commuting observables leading to fundamental uncertainties in measurement outcomes. These unambiguous consequences of the mathematical formalism, which by itself is clear cut, will, as we will show, pose quite formidable epistemological and philosophical questions. It suffices to refer to the infamous Einstein, Podolsky, Rosen (EPR) paradox, which lies at the heart of the well-known Einstein–Bohr debate about how quantum theory defines what we call 'reality.' This debate has been going on for three quarters of a century and only now appears about to be settled.

Going from left to right in Figure 3 is, in some sense, a perspective marked by experimental discoveries and as such a rather historical perspective. Going from left to right is therefore hard because it is erratic, and it moves slowly except for sudden jumps. It is highly unpredictable because it basically lacks internal logic: there is no strictly logical path from classical to quantum physics. The path from left to right is the historic one, and therefore bumpy, but also paved with would-be miracles and intriguing misconceptions, which indeed make a wonderful narrative with ample heroism and drama.

But once the subject has matured, there is the other possibility, namely to start on the right with the concepts and a logical, abstract framework, and from there move back to the left. A theorist like myself naturally prefers a presentation from right to left, which in a sense is highly anti-

chronological, but would be more comprehensive because it has an internal logic and systematics. I believe that once things are understood, going from right to left is *easy*. Moreover, it would give the author the freedom to limit himself to the *quantessence* of a coherent body of knowledge.

Yet, in spite of this argument, it would be a bad idea to really treat the three layers sequentially from right to left, because you need the stuff on the left to appreciate the content of the right column. This suggests the option of a left-right compromise, or left-right coalition, just like that is often the case in the politics of healthy democracies.

Combining parts and layers: the outline. After some reflections on the general structure of the book, let me now just give a more detailed description of the layout of the chapters. As mentioned, I have divided the book in three parts or volumes, as indicated in Figure 2. Volumes I and III are primarily descriptive and do not require much math, since they are phenomenologically-oriented. So in the context of the layers, Volumes I and III mainly deals with $\mathcal{A}$ with some attention to layer $\mathcal{B}$. Volume II, with a title that refers to the 'quantessence', focusses more on the mathematical and conceptual structure of the theory, and covers the layers $\mathcal{B}$ and $\mathcal{C}$. As a matter of fact, quantum lovers with an outspoken fear of formulas may prefer to read only Volumes I and III as a single coherent descriptive account of what quantum theory has achieved. The following preview may help you to make up your mind.

Volume I talks about *The journey*, where we follow a path starting at the level of atoms, and descending deeper into matter to the worlds of nuclei and elementary particles and their interactions. This part is so to speak *inward bound*. But before we embark on this descent in Chapter I.4, we give a review of what classical physics is about in Chapter I.1. Chapter I.2 deals with the very breakdown of classical physics, from which crises the theories of relativity and quantum emerged. Here we also included a section on the physics of geometry and a section on the notions of information and computation, highlighting another fundamental turning point in twentieth century science and technology. In Chapter I.3, on units, scales and universal constants, we obtain surprisingly deep insights in the domains of validity of our cherished theories by applying what we call 'dimensional analysis.' It provides us with a heuristic quantitative sense of what the characteristic scales in nature are, and why. In Chapter I.4 we describe the quest for the basic building blocks of matter all the way from atoms down to the most fundamental constituents of matter and radiation.

In Volume II – called *Quantessence: how quantum theory works* – we give an accessible introduction to the mathematical modelling tools and representations that comprise quantum theory, including those which led to a number of remarkable conceptual and semantic puzzles. This part emphasizes the two deeper layers I alluded to before, i.e. the layers $\mathcal{B}$ and $\mathcal{C}$ of Figure 3.

This second part also leads us deeper into the subjects of quantum information and computing. To that end we first have to contrast the setting of quantum theory with its classical counterpart. In Chapter II.1, the first of Volume II, we start by introducing quantum states as vectors in Hilbert space. I discuss the structure of Hilbert space for qubits and quantum information in quite a lot of detail. In the second Chapter II.2, I discuss the quantessence of observables, why we think of them as operators acting on Hilbert space, and what it means to make a quantum measurement. In this chapter the Heisenberg uncertainty relations are also introduced. In Chapter II.3 I talk about the measurement process more extensively with a particular focus on quantum interference phenomena. Chapter II.4 examines quantum entanglement and some of the modern experiments addressing the profound questions of cloning, Schrödinger's cat, hidden variables, as well as quantum teleportation and computation. In Chapter II.5 I explain the concepts of quantum particles, fields

and strings. There, the famous equations of Schrödinger, Heisenberg and Dirac that describe the time evolution of states and observables, will be introduced. I also explain properties like quantum spin, quantum statistics and their relationships. In Chapter II.6, the final chapter of Volume II, we introduce the notions of symmetry and symmetry breaking which play a central role in all of modern physics. The notion of symmetry served as a powerful guiding principle in our quest to understand nature.

In the third and final Volume of the series we return to model physical reality but we now move upwards from the atomic scale. In Chapter III.1 we discuss how matter sequentially evolved in the very early universe, from quarks, to nucleons, to atoms, and from simple molecules to the basic (bio)chemistry concerning the molecules of life.

Chapter III.2 and III.3 are devoted to the splendid diversity of quantum phenomena in the physics of many body systems that are manifest in gaseous, liquid as well as solid phases. Where in Chapter III.2 we consider the atomic and nuclear lattices and to what extent these are ordered, the focus in chapter III.3 is on the electronic behavior in solids and the quantum phenomena they display.

In Chapter III.4, we touch upon the quite advanced notions of scale dependence and renormalization. Part III could well be called *outward bound*, certainly if reasoned from the atomic scale where quantum theory made its first appearance. The criteria of inward and outward bound refer to the arrows in the left column of Figure 2.

In the concluding Chapter III.5, we zoom out and look at the meaning and impact of quantum in the broader context of science, technology and society.

After the concluding chapter you find a set of *Math Excursions,* appendices in which we offer rather minimal but tailor-made introductions to the mathematics used throughout the book.

Choosing the structure of a three-volume book means that we couple together the layers of Figure 3 so as to enable a coherent presentation of the quantessence as a whole, which is accessible without being too superficial. What you see is that quantum theory, even when you restrict yourself to the *quantessence*, is a huge field. and that is why I divided the book up in three volumes.

As you may be aware, an impressive number of Nobel prizes have been awarded in the course of the past century to quantum discoveries in physics and chemistry. We list most of them in an appendix (on page 644) at the end of the final volume. There we also provide some of the chronology, and list the names of many of the influential thinkers who made seminal contributions to the field. It may also help you to follow up specific topics that have caught your interest while reading.

I like to think of the three parts of the book as a kind of a triptych, where the central panel covers the deeper quantum scenery, while the side panels are more descriptive and discuss lots of real physics, from quarks all the way up to bio-chemistry and the splendid diversity we encounter in the condensed states of matter.

SANDER BAIS

The journey: from classical to quantum worlds

The Quantessence of Reality

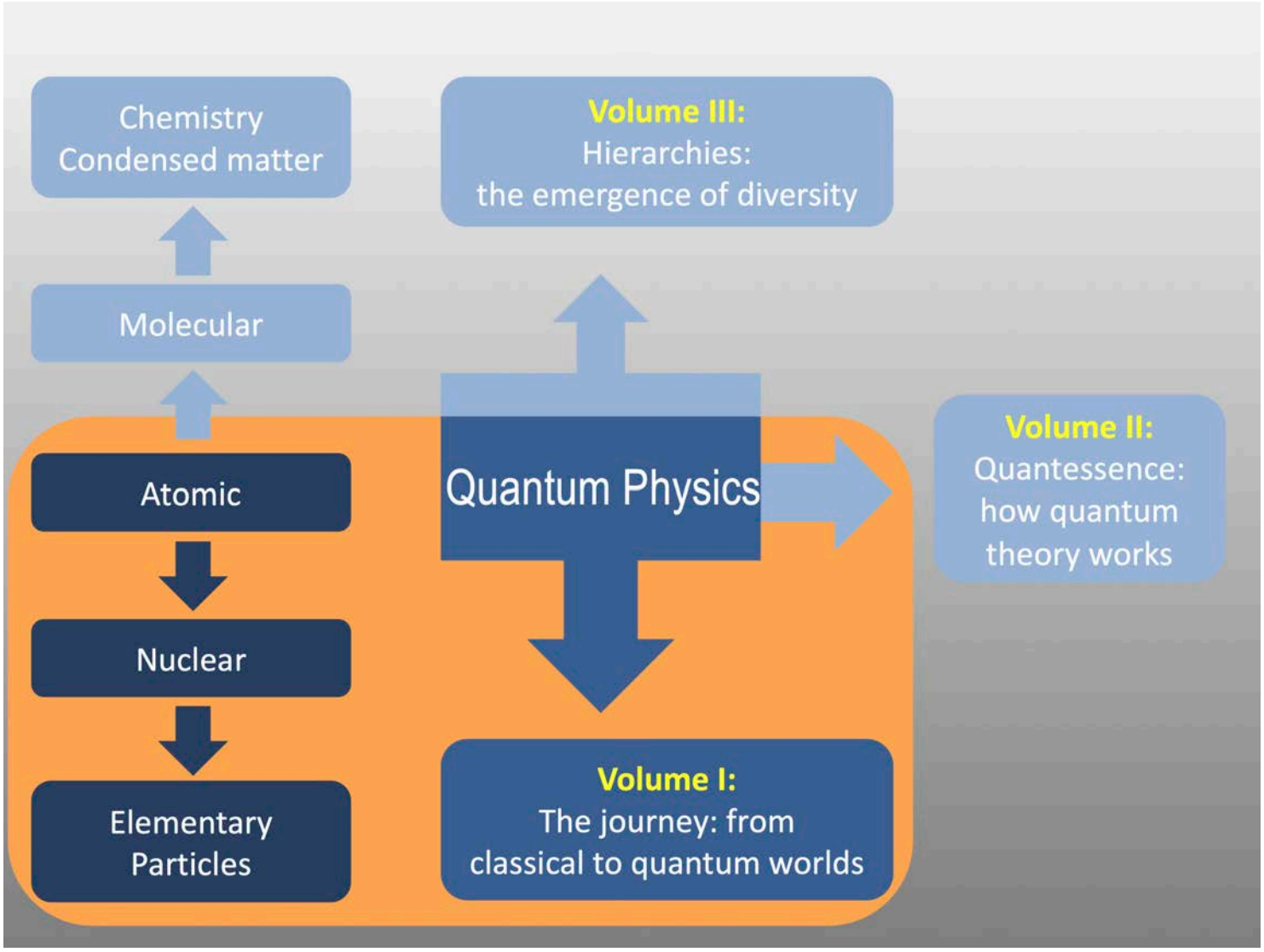
Amsterdam
University
Press

The Journey: From Classical to Quantum Worlds

We start this volume by a brief explanation of the core results of classical physics and how observations necessarily led to the turning points of quantum and relativity. There is a chapter on the age of geometry and information and one on the fundamental constants of nature and their meaning, where we to a certain extent we cover Big Bang cosmology and black holes. We then describe the magnificent race to the bottom of discovering ever more elementary layers of particles and the fundamental forces between them culminating in the Standard Model. More speculative subjects like grand unification, superstring theory and the multiverse are finally also touched upon.

Volume I

The journey:
from classical to quantum worlds


Chemistry
Condensed matter
Molecular
Atomic
Nuclear
Elementary
Particles
Quantum Physics
Volume III:
Hierarchies:
the emergence of diversity
Volume II:
Quantessence:
how quantum
theory works
Volume I:
The journey: from
classical to quantum worlds

Chapter I.1

The gems of classical physics

Mission almost completed

> Don't laugh! There is a special section in purgatory for professors of quantum theory, where they will be obliged to listen to lectures on classical physics ten hours every day.'
>
> Paul Ehrenfest (in a 1927 note to Einstein)

In this chapter we recall the great theories of classical physics with an emphasis on their underlying principles. You could call them 'classessential.' One by one they represented turning points in our understanding of Nature. First there are the four fundamental laws of Newtonian mechanics and gravity, which open the door tto the precise description and modeling of general dynamical systems through systems of differential equations. The second pillar corresponds to Maxwell's four laws of electromagnetism, describing all classical electromagnetic phenomena including those of light and radiation. The third component consists of the macroscopic laws of thermodynamics and how they are explained from the underlying microscopic dynamics by the theory of statistical mechanics. It is here that the notion of entropy emerges as a measure of information or disorder.

Towards the end of the nineteenth century, physics almost came to an end. The physics community seemed optimistic and self-confident. Most of the observed natural phenomena appeared to be accounted for in the established framework of classical physics, at least in principle. It was hard to imagine what else could present itself to their curious minds. And indeed, after many centuries of scientific research, there had been not only highly impressive technical achievements, but fundamental laws of nature had also been established.

I like to warn agains the sloppy use of terminology when it comes to words like 'models,' 'theories' and even 'laws of nature.' It is rule rather than exception among scientists to treat these concepts as basically interchangeable. We refer to the theories of Newton and Einstein, by which we mean the mathematical description of certain theoretical models of nature's behavior, which in fact have been so successful that they are often also referred to as 'laws of

nature.' Yet they are not true in any absolute sense, there is no guarantee that we will not one day find that nature violates such a law of nature in a domain of reality that we cannot yet observe. So, on the one hand, one might as well put these cherished 'laws' in the category of 'working hypotheses' in view of the fact that we can never fully exclude the possibility that they are conceivably false. On the other hand, the 'laws' have proven to be remarkable robust quantitative statements on the workings of nature that have survived centuries of ever more extensive (and expensive) experimental tests. In that sense they express some of the core messages carried by nature about our world, about what and who we are, and how things ended up this way. They may not tell us *why* we are here but at least *how* we got here. It appears that modern science in many ways liberates us from the narrow anthropocentric views that are as dominant as they are questionable in the debate of what the place and future of humankind in this universe may be.

When I talk about the breakdown of classical physics, I refer precisely to the type of breakdown where the declared universality of laws turned out to primarily express our overconfidence. The term breakdown here is not as much a matter of whether a theory is right or wrong, but rather marks the limited domain of validity of any particular theory. In any pragmatic sense there is nothing wrong with classical physics as long as you apply it to problems within its domain of validity. You may compare it to the situation in biological evolution where it is evident that we have passed beyond the stage of bacteria, but that doesn't stop them from being around and still playing a crucial role.

What the notion of classical physics refers to depends on the context in which it is discussed. Often 'classical' is contrasted with 'quantum', and in that case we can consider the theory of relativity to be part of classical physics. We could however also contrast 'classical' with 'modern', and in that case we can draw the line at the end of the nineteenth century and count both relativity and quantum theory as parts of 'modern' physics. It is this latter distinction that we will make in this chapter. The use of the word 'modern' will strike you remarkably inappropriate because this 'modern' physics was to a large extent formulated a century ago; 'modern' in this context clearly does not mean contemporary. Modern theory in this context apparently just means that we have not yet encountered the limits of its domain of validity. In this chapter we start by briefly recalling the core messages of the classical theories of mechanics and the gravitational force, of the theory of electromagnetism and light, and of the theories of thermodynamics and statistical physics.

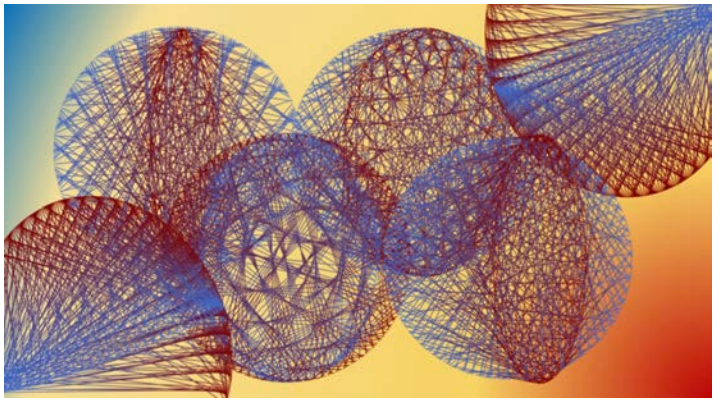


Figure I.1.2: *Newtonia: Composition with bound orbits.* (Image constructed using visualization & graphics tools of the Mathematica package.)

In the next chapter we briefly summarize how certain crises in classical physics seeded two fundamental turning points in our thinking about nature: relativity and quantum physics. In that chapter we also introduce the basic concepts of information theory, as this branch of science is now also heading towards a quantum revolution.

In the third chapter we delve deeper into the notion of the domain of validity of a model and discuss how the particular values of the universal constants that appear as parameters in physical models basically set the scale of our universe.

The fourth chapter gives an account of our progressive insights in what the basic building blocks of nature are, from the atomic level all the way down to superstrings.

Newtonian mechanics and gravity

Newton's work lead to the unified description of terrestrial and heavenly mechanics and involved the creation of the mathematics of change, called differential calculus, which in turn gave rise to the birth of the general theory of dynamical systems.

Four laws only

Back to the achievements of classical physics. Firstly there are Newton's four laws described in his genial *Principia Mathematica* published in 1667. Three of those laws constituted the foundations of mechanics: (i) the law of inertia, (ii) the force law and (iii) the the law of action and reaction. The fourth law is the law that defines the gravitational force between two masses.

The first law: the law of inertia. The law of inertia postulates that if a body is at rest or moving at a constant speed in a straight line, it will remain at rest or keep moving in a straight line at constant speed unless it is acted upon by a force. This property is called *inertia*. We have illustrated the distance traveled $\mathrm{x}(\mathrm{t})$ for a body of some mass m, for two constant velocities $\mathbf{v}_1 < \mathbf{v}_2$ in Figure I.1.3. In the absence of a force the distance traveled is proportional to the elapsed time, in other words: $\mathrm{x}(\mathrm{t}) = \mathbf{v}\mathrm{t}$.[1] The *first law* led

[1] We adopt the notational convention where symbols denoting vector-like quantities like position, velocity, momentum and force are given in bold except when we are dealing with one spatial dimension. For the length and the length squared of a vector we write $|\mathbf{v}| \equiv v$ and $\mathbf{v} \cdot \mathbf{v} \equiv |\mathbf{v}|^2 \equiv v^2$. Scalar quantities like mass are set in the default typeface.

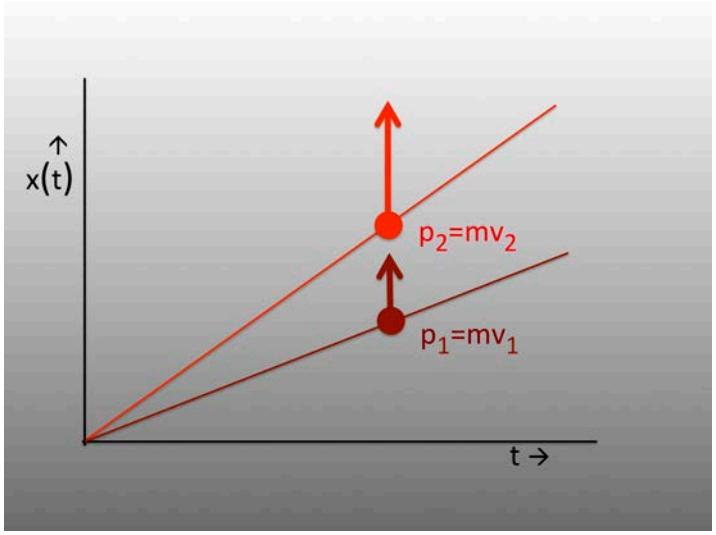


Figure I.1.3: *Newton's first law.* In the absence of a force a body will move at a constant velocity and momentum. In the figure the distance traveled as a function of time $\mathrm{x}(\mathrm{t})$ for a body of mass is plotted for two constant momenta p_1 and p_2, corresponding to the two arrows.

to the fundamental notion of momentum, where the *momentum* $\mathbf{p}$ of an object is defined as the product of its mass m and its velocity, $\mathbf{p} = \mathrm{m}\mathbf{v}$. This linear relation between momentum and velocity is depicted in Figure I.1.4 , where the slope of the line by definition equals the mass. Momentum is also referred to as the 'amount of motion,' and if you don't have a feeling for it, think of it as impact. If somebody throws a large brick to you the impact will be much larger than when that same person would have thrown a piece of foam of the same shape with the same velocity. The first law states that in the absence of a net force on an object its momentum will not change. Zero force means that momentum is conserved, and this implies that the velocity is constant.

The second law: the force law. The *second law*, called the *force law*, is the well-known relation between the force $\mathbf{F}$ applied to a body, and the resulting acceleration $\mathbf{a}$, given

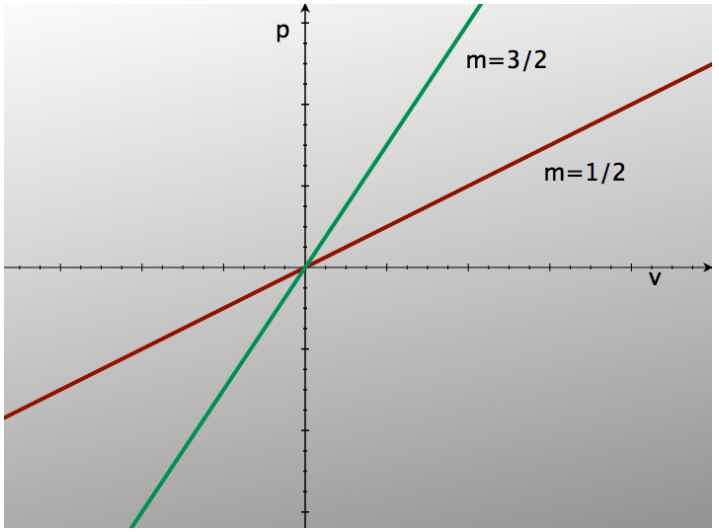


Figure I.1.4: *Definition of momentum.* Newton defined $\mathrm{momentum}$ as the 'quantity of motion' directly proportional to the velocity of the object, the proportionality constant equals the mass m of the object.

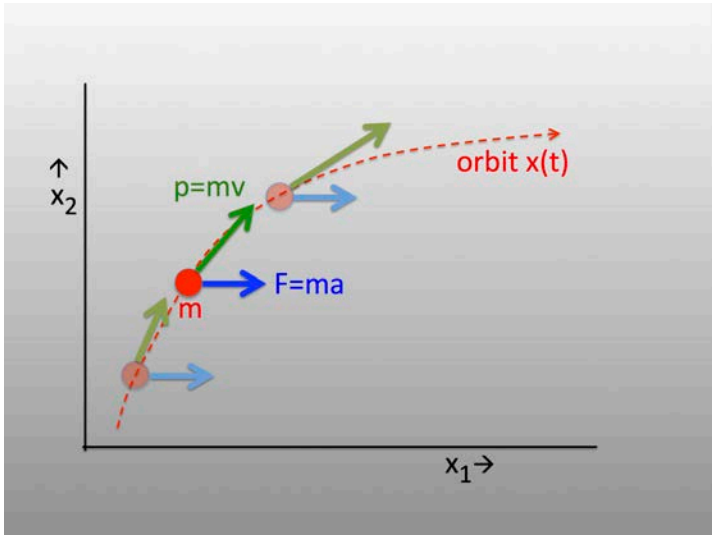


Figure I.1.5: *Newton's second law.* We have drawn a segment of the orbit in 2 dimensions of a particle with mass m under a constant force $\mathbf{F}$ in the x_1-direction. This could be a particle with charge q in a constant electric field $\mathbf{E}$ exerting a force $\mathbf{F} = \mathrm{q}\mathbf{E}$. Note that the momentum $\mathbf{p}$ at time t points along the slope of the particle's trajectory $\mathbf{x}(\mathrm{t})$. Because the force $\mathbf{F}$ points in the x_1-direction, only the component p_1- will increase while p_2 remains constant.

by the fomula $\mathbf{F} = \mathrm{m}\mathbf{a}$.[2] As acceleration is the rate of change in velocity, the force is then equal to the rate of change in momentum. A brilliant aspect of this equation is that at first glance it doesn't seem to hold. I remember as a kid pulling other kids on a sled through the snow: yes I had to pull to get the sled moving but if I stopped pulling it did not keep moving with constant velocity as I thought should be concluded from the law. No force, no change in momentum. But the sled immediately came to a halt after I stopped pulling it. I had to conclude that there should be another force in action, and indeed there was, it was the resistive force of the snow. Now that is a funny force that opposes motion, the greater the velocity, the greater the force in the opposite direction. It is as subtle as the workings of the opposition in parliament. But postulating a force with such a subtle adaptive power, is that not just postulating what you see, postulating the facts you wished to explain? Well don't put the book aside yet, there is more to come.

The third law: action is reaction. A simple example of the law of 'action is reaction' is provided by a book at rest on a table as depicted on the left in Figure I.1.6. Gravity pulls the book down (light blue arrow) attached to centre of mass pointing down), it equals the force of the book on the table (dark blue arrow pointing down) and indeed, the book would fall down were it not for the table exerting an equal but upward directed *normal* force on the book It is this balance of forces that act on an object, which is the main topic of *statics*. It means that the net force, but also the net torque, on an object should be zero and that does

[2] Actually it should be written as $\mathbf{F} = \mathrm{d}\mathbf{p}/\mathrm{dt}$, where strictly speaking there is an extra contribution because $\mathrm{d}\mathbf{p}/\mathrm{dt} = \mathbf{v}\mathrm{dm}/\mathrm{dt} + \mathrm{m}\mathrm{d}\mathbf{v}/\mathrm{dt}$. The first term proportional to the change in mass is considered to be zero because for a single particle one assumes a constant mass. But for a rocket burning its fuel this is no longer true.

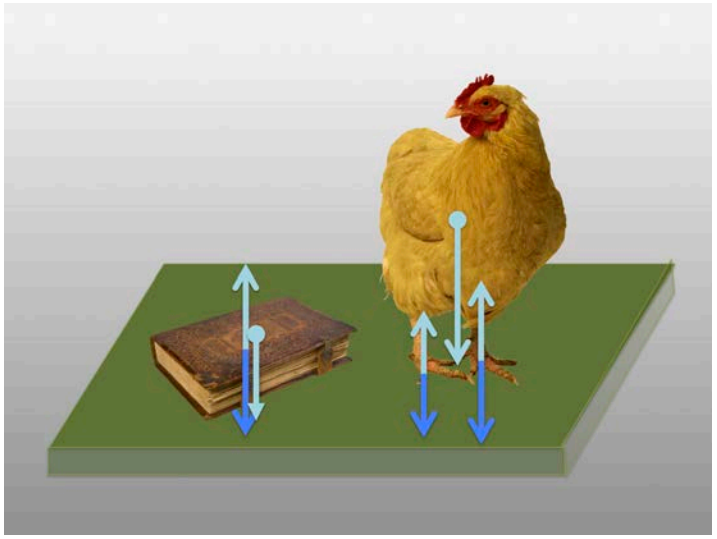


Figure I.1.6: *Newton's third law.* The third law *Action = Reaction* applies to a chicken at rest on a table. The downward gravitational force can be represented by the large light blue arrow attached to its centre of mass. Through its legs it exerts in two places a force on the table, and the table exerts a *reaction force* exactly equal and opposite at the points of contact. The net force, which is the sum of the light and dark blue arrows on the chicken, is zero and its change in momentum will be zero so it doesn't move. But why are the forces of the two legs unequal? That is to make sure that the chicken doesn't fall over sideways. This requires that the *torque* on the chicken has to be zero as well, so that its angular momentum does not change.

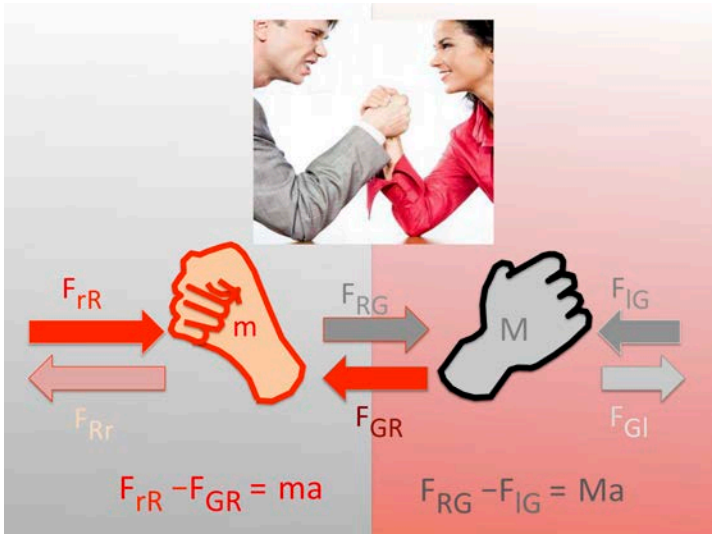


Figure I.1.7: *Newton's third law.* The third law *Action = Reaction* applies to arm wrestling, also when the balance of power is broken and the 'red' force is larger than the 'grey' force. An explanation is given in the text.

not only explain the stability of architectural structures like bridges, arches or cathedrals, but also the stability of the chicken at rest on the table at the right-hand side of the figure.

A more subtle example of the *third law* is provided by a game called 'arm wrestling.' Two individuals (still mostly men) sit at opposite sides of the table and fix their elbows on the table and try to push each other's hand towards the table. For quite some time nothing seems to happen, in spite of the fact that both individuals do their utmost best to get the fists moving. As long as nothing happens the net forces of the hands are in perfect balance, a situation that is called a static equilibrium. This lasts until the balance is broken, leading to a net force in one direction causing both fists to start moving, until one hand is forced on the table and somebody has to order a round of beer.

The question you may wrestle with is whether in such a dynamic situation the action-is-reaction-law still holds. So, let us look more closely at how to apply the fundamental action-is-reaction-law in such a dynamic setting. This is explained in Figure I.1.7, where we give a schematic of the forces involved. We identify three different instances where the third law can be applied. Firstly, on the left side we have the force of the red arm r on the 'red' hand R (denoted by $\mathrm{F_{rR}}$), which indeed equals the opposite force of the red hand on the red arm ($\mathrm{F_{Rr}}$). In the middle we have the force of the red hand on the grey hand ($\mathrm{F_{RG}}$), and the force of the grey on the red hand ($\mathrm{F_{GR}}$), these have to be equal because of the third law applied at the interface between the hands. On the right side we have the force of the grey arm ($\mathrm{F_{lG}}$) on the grey hand and the equal and oppo-

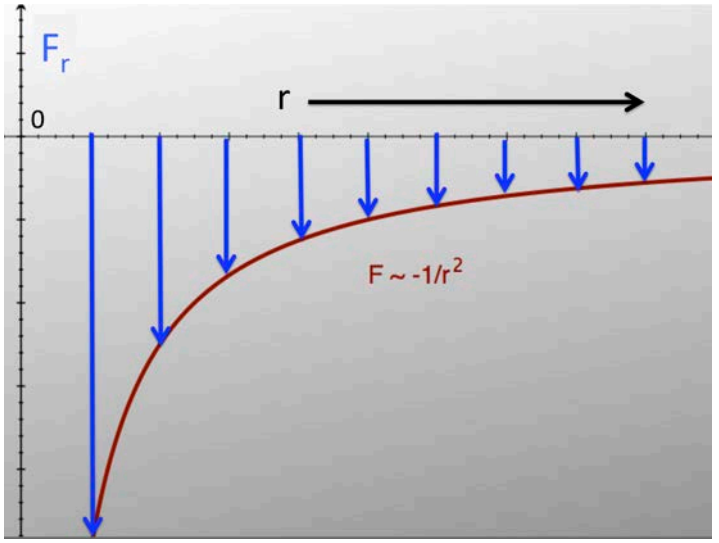


Figure I.1.8: *Newton's fourth law.* This is Newton's famous 'inverse square law' for the gravitational force between two massive objects as a function of their distance r. The force only has a radial component, which is negative meaning that the force is attractive.

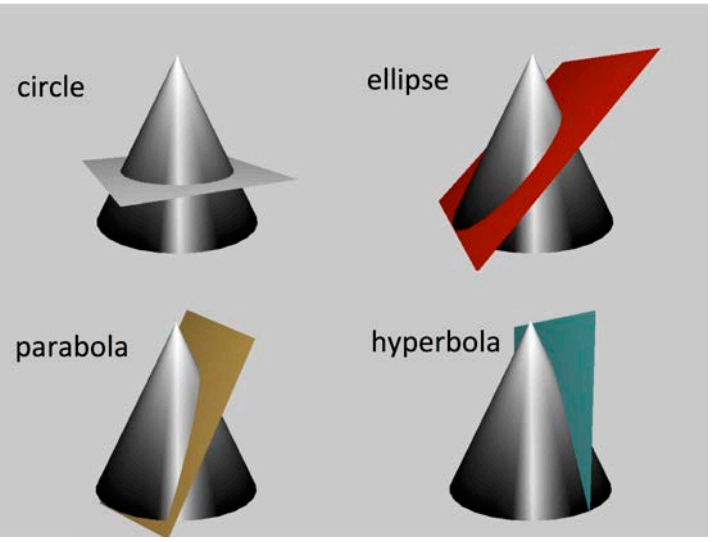


Figure I.1.9: *Conic sections.* The general solution for the orbits of a planetary object around a star can be obtained by inserting the gravitational force in the second law. The resulting orbits correspond to the conic sections depicted above.

site force (F_{Gl}). So we see that the third law should be applied three times referring to three different forces. If both hands move with an acceleration a we can firstly apply the force law to the red hand telling us that the net force on it is $F_{rR} - F_{GR} = ma$, applying it to the grey hand it yields $F_{RG} - F_{lG} = Ma$. Next we use the result that $F_{RG} = F_{GR}$ to ascertain that the net force $F_{rR} - F_{lG} = (m + M)a$, which is the force law applied to the system of both hands. This argument shows that the hands can be in accelerated motion, not in spite of, but rather thanks to the fact that the law of 'action is reaction' remains valid all along. It illustrates the important fact that 'action is reaction' is a general law, that is applicable as long as the objects exerting force on each other stay in contact.

The fourth law: the law of gravitation. Newton's *fourth law* is his celebrated *universal law of gravitation*,

$$F_r = -\frac{G_N m_1 m_2}{r^2}, \qquad \text{(I.1.1)}$$

expressing the attractive gravitational force between two masses m_1 and m_2 as proportional to the inverse square of their distance r. The force as a function of the distance, is depicted in Figure I.1.8. Note that here the principle of 'action is reaction' is indeed respected implicitly, because it is the force 'between' two objects, they experience an equal force in opposite directions. Indeed it is so universal that it applies with the same constant G_N equally well to a pencil dropping on the floor (the earth) as to the motions in the solar system or to the motion of stars in the Milky Way. It was justly said that Newton with this law unified celestial and terrestrial mechanics. Substituting this gravitational force in the second law, one can solve the system for general planetary orbits around a star. They correspond to the well-known *conic sections* depicted in Figure I.1.9, where the top two are the bound circular or elliptic orbits, and the bottom two are the unbound parabolic and hyperbolic orbits.

Dynamical systems

Let me say a little more on how these laws of Newtonian mechanics furnished a first and powerful description of what nowadays is called a *dynamical system*, a system described by a set of variables whose values change over time. Thinking about mechanics that way, one would rewrite the laws in a different way that illuminates the dynamical system's perspective.

Phase space. First we say that if we look at a particle as a system, then it has at any instant in time a state that is labeled by two variables, its position x and its momentum p. So, we may think of the state of the system as corresponding to a point in (x, p)-space. This space is usually called the *phase space* $\mathcal{P}_{ph}$ of the system. For a particle moving in ordinary space $\mathcal{P}_{ph}$ is six-dimensional, because we have to specify the three components of its position and the three components of its momentum. The dynamics of the system can be envisaged as a trajectory $(x(t), p(t))$ of the point that represents the state of the system, through $\mathcal{P}_{ph}$. This trajectory is then specified by giving the rule which tells you where the system goes if you give the point at some initial time t_0.

Differential equations. This rule is like an incremental prescription, it specifies an infinitesimal change by using the notion of a (time) *derivative* (d/dt) as a measure of change:

$$\frac{d\,\mathbf{Something}}{dt}\Big|_{t_0} = \begin{cases} change\ of\ that\ '\mathbf{Something}' \\ per\ unit\ time\ at\ t = t_0. \end{cases}$$

Equations involving this (time) differential are called *differential equations*, to contrast them with algebraic equations – like the quadratic equation $ax^2 + bx + c = 0$ – in which algebraic expressions in the variables appear but no derivatives. If the equations involve time derivatives, we speak of the *equations of motion*. If the system is *closed*, the change will depend only on the state of the system at earlier times. In the quite common case that the system has no memory – like the system of the sun and the earth in the Newtonian picture – the change at some particular time t only depends on the state of the system at time t. It is generally agreed upon that sun and earth do not wrestle with sleepless nights caused by bad memories. So the dynamical system with a set of N independent variables $\{y_i\}$ with $i = 1, \ldots, N$ would look like a set of N coupled equations describing the change of the system by specifying the N components f_i of the change vector, each of which may in turn depend on the set of all variables $\{y_j\}$:

$$\frac{dy_i}{dt} = f_i(\{y_j\})\,. \tag{I.1.2}$$

The functions $f_i(y_j)$ encode the interactions between the different variables, In other words, these variables include their mutual dependence and of course a number of external parameters which typically appear as the coefficients

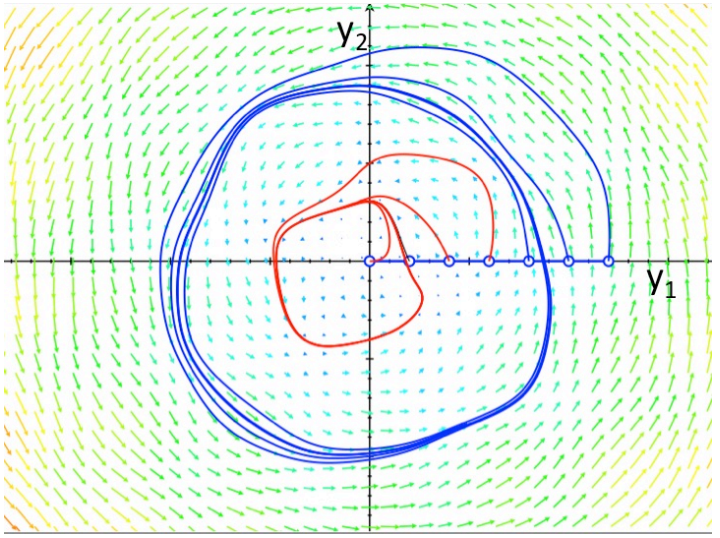


Figure I.1.10: *Dynamical system.* We display the vector field corresponding to a particular example of (I.1.2). This means that in each point (y_1, y_2) we plot the vector (arrow) with components $dy_1/dt = -y_2 + \cos 2y_1$ and $dy_2/dt = y_1 + \sin 2y_2$. Solutions of the system correspond to trajectories that start from a given point in (y_1, y_2)-space, following the arrows. In this case we give some trajectories starting on the y_1 axis that converge either to the blue or the red limit cycle.

of the terms specifying the interactions. The $f_i(y_j)$ correspond to the components of the 'change vector' at any point in phase space, which means that they define a *vector field* over the configuration or $(\{y_i\})$ space. This vector field forms a powerful mathematical representation of the dynamical system as a whole. It depicts the phase space as a fluid flow. If we drop autumn leaves into the flow, they will start to move, following the particular flow lines which correspond to the particular solutions of the dynamical system. We have depicted a particular vector field in Figure I.1.10 , which also shows two sets of trajectories described by solutions of the dynamical system for different starting points on the y_1-axis. The trajectories are obtained by locally following the direction of the vector field. The dictum is indeed: 'go with the flow.' The orbits are seen to converge on one of two different closed limit cycles.

Yet another way to look at a dynamical system is that it represents an algorithm that takes input information, the vector defining the initial point in phase space and moves or 'processes' it, to some final state.

Writing Newtonian mechanics in this format the first and second laws look like,

$$\begin{aligned} \frac{d\mathbf{x}}{dt} &= \mathbf{p}/m\,, \\ \frac{d\mathbf{p}}{dt} &= \mathbf{F}\,. \end{aligned} \qquad \text{(I.1.3)}$$

They completely specify the motion of the point in phase space, where the force $\mathbf{F} = \mathbf{F}(\mathbf{x}, \mathbf{p})$ may in general depend on the position and velocity of the particle. It is customary to treat the earth-sun system by keeping the sun fixed in the origin (a good approximation because the sun has a huge mass) and let the earth move through the gravitational force (the fourth law) that only depends on the length of the position vector $r = |\mathbf{x}|$. The third law is basically a *constraint* on the system: if we had included the position and momentum of the sun as independent variables, then the third law would require that the same force F would appear with the opposite sign in the equations for the sun and for the earth respectively. From this example it is also clear that the functions on the right-hand side of the equations do not only depend on the variables, but also on certain parameters that set the strength of the couplings or interactions. These parameters, like the masses or the Newton's gravitational constant, are supposed to be constant but must of course be varied to find the best fit to the experimental data. They are the input parameters of the model. It is here that *Occam's razor* – the principle of rational minimalism – applies, decreeing that if two models perform equally well, the one with the fewest parameters is to be preferred.

Conservation laws

> The tears of the world are a constant quantity. For each one who begins to weep, somewhere else someone stops. The same is true for laugh.
>
> *Samuel Becket – Waiting for Godot*

Note that with the dynamical laws for the fundamental variables, one can also calculate the time evolution of other ($\mathbf{x}-$ and $\mathbf{p}-$dependent) dynamical variables. One such variable is the energy, often called the Hamiltonian and denoted as H . It should be thought of as a function $H = H(\mathbf{x}, \mathbf{p})$ of the basic state variables $\mathbf{x}$ and $\mathbf{p}$. Another such variable is the angular momentum $\mathbf{L} = \mathbf{x} \times \mathbf{p}$, which is basically the amount of rotational motion, or the rotational momentum. We will return to these quantities shortly.

Under certain circumstances it may happen that some dynamical variables are *conserved*, meaning that they do not change over time. These are often called *constants of the motion.* For example in Newtonian mechanics, if there is no force, that is we have that $\mathbf{F} = 0$, then the equation (I.1.3) tells us immediately that the momentum does

not change, it stays constant and is thus 'conserved.' On the other hand, if the force only depends on the distance and not on the direction (as is the case in Newton's gravitational force), then the angular momentum will be conserved as we will explain shortly. So, if the time derivative of some physical quantity Q equals zero:

$$\frac{dQ}{dt} = 0\,, \tag{I.1.4}$$

we call the equation a *conservation law* for Q, because the amount of Q is constant in time.

Energy conservation. Of special interest is the case of energy conservation because it is of general validity and applies to basically all observed processes in nature that are physically based. Let us for convenience restrict ourselves to a (one-dimensional) situation which is simple but also surprisingly common, where the energy H consists of two parts, a *kinetic energy* part $U(p)$ which only depends on the momentum, and a *potential energy* part $V(x)$ that only depends on position:

$$H(x,p) = U(p) + V(x)\,. \tag{I.1.5}$$

Then its time derivative can be calculated:[3]

$$\frac{dH}{dt} = \frac{dU}{dt} + \frac{dV}{dt} = \frac{dU}{dp}\frac{dp}{dt} + \frac{dV}{dx}\frac{dx}{dt} = F\frac{dU}{dp} + \frac{p}{m}\frac{dV}{dx}\,.$$

We see that the energy will be conserved if the terms on the right-hand side cancel each other. This requires that the following equalities have to hold:

$$\frac{dU}{dp} = p/m\,, \tag{I.1.6a}$$

$$\frac{dV}{dx} = -F\,. \tag{I.1.6b}$$

[3]The second equal sign involves the use of a mathematical identity called the *chain rule* which says that if $U(p)$ depends on t only through its dependence on $p(t)$, the time change can be found by first calculating the change in U because of a change in momentum, multiplied by the change in time of the momentum. It roughly means that one may cross out the dp factors of the numerator and denominator.

The first condition leads to the well-known expression $U = p^2/2m$ while the second restricts the force in that it has to be equal to minus the spatial change of some potential energy function V. Such a force field is not surprisingly called *conservative*, exactly because its action 'conserves' the total energy of the system. Whereas a 'conservative force' is standard physics jargon, I have never come across terms such as 'liberal' or 'progressive' forces, though if we get to the strong nuclear force, other evocative terms will surface, like 'asymptotic freedom' and 'infrared slavery.'

Applying a (net) force means doing work. If we apply a force F to a mass m, the mass will accelerate and over time its kinetic energy will change. If we push a stroller, we do work by applying a force on it. If we put a charge in an electric field, the charge will start moving because it is the field that exerts a force that causes the motion and it is the field that does the work. The change in kinetic energy ΔU by definition equals the amount of work ΔW that the force has done. If the force is constant, this means that $\Delta W = F \cdot \Delta \mathbf{x}$. If the mass moves through a conservative force field $F(\mathbf{x})$ and it moves along a certain path γ from $\mathbf{x}_0$ to $\mathbf{x}_1$, we know from conservation of the total energy that $\Delta E = 0$, and thus $\Delta U = -\Delta V = V(\mathbf{x}_0) - V(\mathbf{x}_1)$. The amount of work in an arbitrary force field can be expressed as the *line integral* of the force field along a path of motion γ :

$$W = \int_{x_0}^{x_1} \mathbf{F}(\mathbf{x}) \cdot \mathbf{dl}\,,$$

where the line element $\mathbf{dl}$ is the infinitesimal vector tangent to the path γ in the point $\mathbf{x}$. For a conservative force field we get,

$$W = -\int_{x_0}^{x_1} \boldsymbol{\nabla} V(\mathbf{x}) \cdot d\mathbf{x} = V(\mathbf{x}_0) - V(\mathbf{x}_1)\,,$$

and we see that the change in potential energy equals the difference of the potential energies at the endpoints of the path, consistent with the conservation of total energy E. The fact that the difference only depends on the endpoints means that the increase of energy is *not* dependent on the

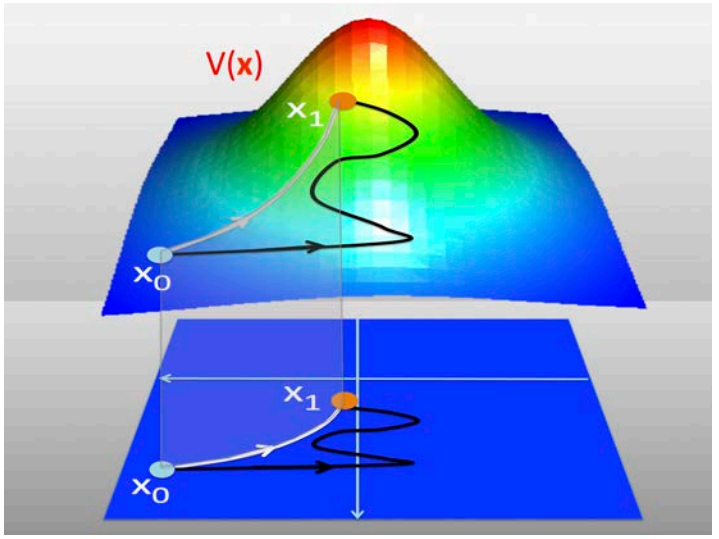


Figure I.1.11: *A line integral.* In the upper picture we give a two-dimensional potential surface $V(\mathbf{x})$. The force field is defined as $\mathbf{F}(\mathbf{x}) = -\boldsymbol{\nabla} V(\mathbf{x})$. If we choose a path from point $\mathbf{x}_0$ to $\mathbf{x}_1$, we can integrate $\mathbf{F}$ along that path. This means that we need to integrate the component that is tangential to the path. This line integral yields the value $W = V(\mathbf{x}_0) - V(\mathbf{x}_1)$ which equals the work performed by the force, which in this case is negative. We had to perform a force to go uphill and therefore the potential energy was increased. Note that the outcome is *independent* of the path chosen.

particular path chosen. If you want to climb to the top of a mountain, you can choose between a path that is long and not so steep or a very short, very steep path, in either case you would have to deliver the same amount of work.

The harmonic oscillator. A simple example of a conservative force is the one-dimensional elastic force, applied to a mass m hanging on a spring attached to a beam as depicted in Figure I.1.12,

$$F = -kx\,, \tag{I.1.7}$$

where x is the deviation of the mass from its equilibrium position, k is the elastic constant that characterizes the spring and the minus sign indicates that the force the string exerts is opposite to the displacement. The force tends to

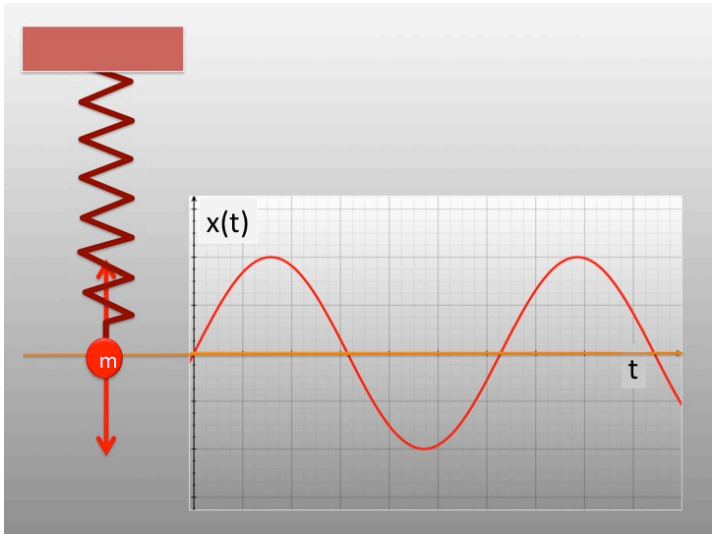


Figure I.1.12: *The oscillating mass .* A model system consisting of a mass m attached to a spring. The inset shows the oscillatory motion of the mass in configuration (x,t)-space.

restore the equilibrium state. Because the force increases linearly with the distance x, and according to the equation (I.1.6) it has to equal minus the derivative of the potential energy, we may conclude that the corresponding potential V satisfying that condition has to grow quadratically with x (up to an irrelevant constant term):

$$V(x) = \frac{1}{2}kx^2\,. \tag{I.1.8}$$

We have depicted the energies V, U and H corresponding to the resulting oscillatory motion in Figure I.1.13. The spring keeps oscillating with a fixed frequency, which is equal to $\sqrt{k/m}$, a fixed amplitude and a fixed total energy H. These motions correspond to the configuration space picture of Figure I.1.12, and phase space picture of Figure I.1.14. This *harmonic oscillator* is quite ubiquitous, because systems are most of the time in equilibrium. And if we perturb such a system, it typically starts oscillating around its equilibrium configuration and in real cases it usually relaxes back to equilibrium because of frictional forces. So the harmonic potential is the simplest approximation that corresponds to the 'linear' response of

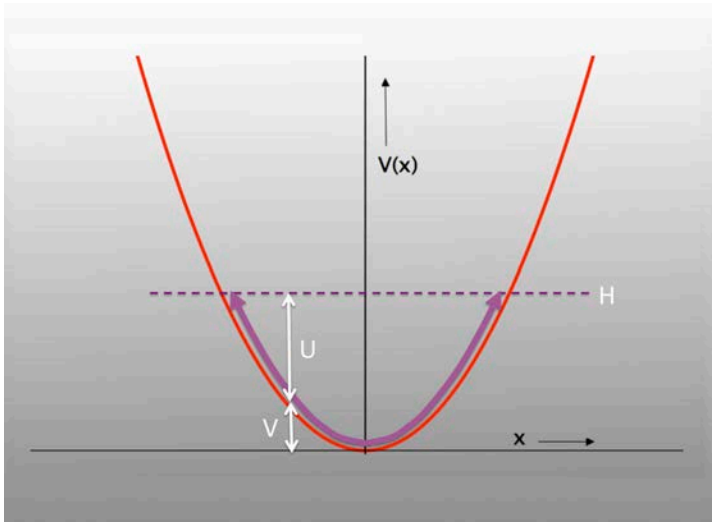


Figure I.1.13: *The harmonic oscillator.* The harmonic potential $V(x) = -\frac{1}{2}kx^2$ in red. The equilibrium point is at the origin, the resulting force is $F = -kx$ and is always directed towards the equilibrium point. If there is no friction, the position x will oscillate around the origin with a fixed amplitude and a fixed total energy H.

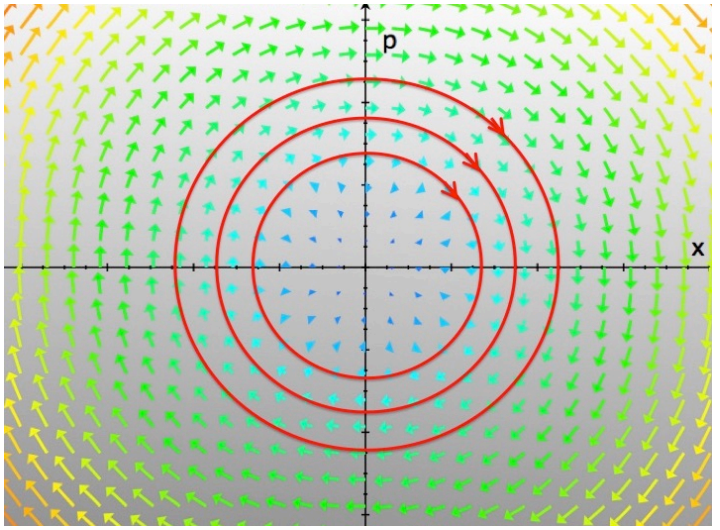


Figure I.1.14: *Periodic orbits.* The phase space vector field corresponding to the harmonic oscillator with $m = k = 1$ becomes $(dx/dt, dp/dt) = (p, -x)$. The orbits correspond to limit cycles. The origin is a fixed point that coincides with the particle at rest.

the system, which should hold as long as the perturbations are small. This quadratic potential is crucial and will also show up in many different guises at all levels of (quantum) mechanics.

Newton's gravitational potential. The most well-known potential is the gravitational potential due to a mass M located at the origin in Newton's theory, defined as:

$$\mathcal{V}(r) = -\frac{G_N M}{r}\,, \tag{I.1.9}$$

where we are now in three dimensions and r denotes the length of the position vector $r = |\mathbf{x}|$ Note that the potential energy is taken to be zero at infinity. The potential energy of a mass m at a position at a distance r equals $V = m\mathcal{V}(r)$. And it does indeed lead to Newton's celebrated 'inverse square' law (I.1.1). If we let the particle go at some position r, it will move radially inward thereby lowering the potential energy, but at the same rate increasing its kinetic energy so that the total energy remains the same. The conclusion of this part of the story is that if the conditions (I.1.6) are met, the total energy will be conserved if the system evolves according to Newton's laws.

Angular momentum. Another important conserved quantity (in a problem with spherical symmetry) is the angular momentum $\mathbf{L}$, which is a vector quantity just like position or momentum (velocity) and has three components, each of which is conserved. You experience that conservation law if you are cycling. If the wheels spin fast, the angular momentum vector will be directed perpendicular through the axes of the wheels, and the conservation law is reflected in the stability of the bike at high speed. Kids apparently know about this law because they like to take both hands off the handlebars. However, if they slow down, they have to be careful to not tripple over sideways, as small external disturbances may cause a torque that changes the angular momentum, breaking the conservation law.

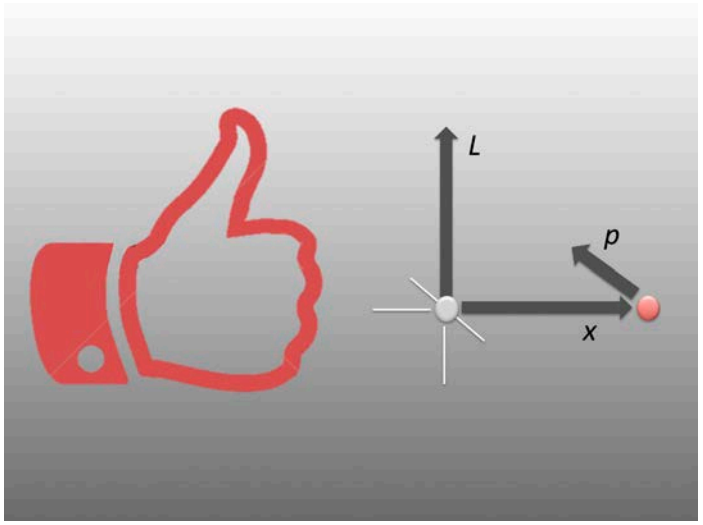


Figure I.1.15: *The Like-rule.* The defining relation of 'angular momentum,' or the 'amount of rotation' of a particle in some orbit. It is given by the vector $\mathbf{L} = \mathbf{x} \times \mathbf{p}$ where the 'times' symbol $\times$ is called the vector product, which is a well-defined multiplication rule for three-dimensional vectors. Whether you like it or not, the Facebook inspired 'Like' symbol on the left symbolizes the 'Like-rule' that tells you in which direction the resulting $\mathbf{L}$ vector is pointing. The instruction is also called the right-hand or corkscrew rule and their importance derives from the fact that they unambiguously link a direction to a rotation.

We have illustrated the defining relation $\mathbf{L} = \mathbf{x} \times \mathbf{p}$ in Figure I.1.15. So $\mathbf{L}$ is a vector perpendicular to the surface spanned by the vectors $\mathbf{x}$ and $\mathbf{p}$. Whether it is pointing up or down is determined by the right-hand rule, which in modern parlance could be better termed the right 'Like' or 'L' rule: point your right-index in the direction of the first vector $\mathbf{x}$, bend your fingers in the direction of the second vector p, then the resulting vector $\mathbf{L}$ will point in the direction of your thumb. This rule explains the meaning of the *vector* or *cross product* or $\times$ sign for vectors. The length of $\mathbf{L}$ is given by the product of lengths of $\mathbf{x}$ and $\mathbf{p}$ times the $sine$ of the angle between them, implying that

$$|\mathbf{x}\times\mathbf{p}| = |\mathbf{x}||\mathbf{p}|\sin\theta = \begin{cases} 0 & \text{if } \mathbf{x} \text{ and } \mathbf{p} \text{ parallel} \\ |\mathbf{x}||\mathbf{p}| & \text{if } \mathbf{x} \text{ and } \mathbf{p} \text{ perpendicular}. \end{cases} \tag{I.1.10}$$

The vector product of two vectors is a vector that is pointing perpendicular to the plane defined by the two vectors, and indeed the product better be zero if the vectors are pointing in the same direction, because then they do not even define a plane.

In three dimensions we have two types of products for vectors. The *dot, inner*, or *scalar product*, which maps a pair of vectors into a number, $\mathbf{a}\cdot\mathbf{b} = |\mathbf{a}||\mathbf{b}|\cos\theta$, and the *cross, exterior,* or *vector* product which maps a pair of vectors into another vector. These definitions may at first sight seem contrived, but the opposite is true: all this symbol-mumbo-jumbo is mostly there because it offers notational convenience, efficiency and transparency.

This crash course of high school and first-year classical mechanics underscores once more that Newton laid the foundations of a general approach to dynamical systems irrespective of what they precisely describe. The variables could refer to either mechanics or to fluid- or electrodynamics, but for that matter they could equally well refer to ecology or economics. By creating the language and syntax of dynamical systems, Newton opened a monumental gateway into scientific thinking and modelling. Indeed, we are standing on the shoulders of giants. ■

Classical mechanics for *aficionados*

In this addendum we present two alternative ways in which classical mechanics can be cast. The reason to do so is that these formulations, though more abstract, are relevant if we move into the quantum domain.

Canonical (Hamiltonian) structure. Let us first recast the setting of classical mechanics in a – what is called – *Hamiltonian form*. It is just a matter of reformulating the same physics in a slightly different but convenient mathematical form. First we note that from the alternative form of the equations (I.1.5-I.1.6), we learn that $\mathrm{d}\mathcal{U}/\mathrm{d}\mathrm{p} = \partial\mathrm{H}/\partial\mathrm{p}$

and $dV/dx = \partial H/\partial x$.[4] Now we can write the equations of motion (I.1.3) in their Hamiltonian form as

$$\frac{dx}{dt} = \frac{\partial H}{\partial p}\,, \tag{I.1.11a}$$

$$\frac{dp}{dt} = -\frac{\partial H}{\partial x}. \tag{I.1.11b}$$

This form of the equations is also called *canonical* and the x and p variables are called *canonically conjugate*.

Poisson structure. Having pushed the juggling with derivatives this far, it pays to go yet one step further and add one more element, which will present classical Hamiltonian mechanics in yet another elegant form. This formulation in terms of *Poisson brackets* was much preferred by Paul Dirac as it brings the classical theory tantalizingly close to its quantum descendants. We should first note that for an arbitrary function on phase space $f(x,p)$ we can derive its time evolution as a first-order dynamical system like:

$$\frac{df}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial p}\frac{dp}{dt} = \frac{\partial f}{\partial x}\frac{\partial H}{\partial p} - \frac{\partial f}{\partial p}\frac{\partial H}{\partial x}\,, \tag{I.1.12}$$

where we used the equations (I.1.11). Next we may define the Poisson bracket of two *arbitrary* functions $f(x,p)$ and $g(x,p)$ by

$$\{f,g\}_{pb} \equiv \frac{\partial f}{\partial x}\frac{\partial g}{\partial p} - \frac{\partial g}{\partial x}\frac{\partial f}{\partial p}\,. \tag{I.1.13}$$

It is an expression which is antisymmetric in f and g, as $\{f,g\}_{pb} = -\{g,f\}_{pb}$. With this definition we can write the time derivative of any function on phase space (i.e. any dynamical variable) as the Poisson bracket with the Hamiltonian:

$$\frac{df}{dt} = \{f,H\}_{pb}\,. \tag{I.1.14}$$

We say that the Hamiltonian 'generates' the time evolution of the dynamical variables. For a conserved quantity Q we have by definition that $dQ/dt = 0$, which by the equation above implies that $\{Q,H\}_{pb} = 0$. A trivial instance is the case $Q = H$, where $dH/dt = \{H,H\}_{pb} = 0$ as it should. In this way, we may also observe that the equations

$$\frac{\partial f}{\partial x} = \{f,p\}_{pb} \;\text{ and }\; \frac{\partial f}{\partial p} = -\{f,x\}_{pb}\,, \tag{I.1.15}$$

hold as well. The first one states that the x derivative, i.e. the effect of an infinitesimal translation in x-space on f, is 'generated' by the momentum p. Finally I should also point out the remarkable relation

$$\{x,p\}_{pb} = 1\,. \tag{I.1.16}$$

Variables which satisfy this relation are called *canonically conjugate*. These classical equations involving Poisson brackets have striking quantum lookalikes in the form of *commutators* as we will explain in the second Volume of the book.

Lagrangian formulation of mechanics. There is another formulation of classical physics that is of great importance, particularly if one turns to relativistic systems. When we think of simple particle mechanics, the formulation uses the coordinate $x(t)$ and the velocity $v(t) = dx/dt$ as dynamical variables. The central quantity now is not the energy but rather the *Lagrangian* $L(x,v)$ defined as:

$$L(x,v) = \frac{1}{2}mv^2 - V(x)\,, \tag{I.1.17}$$

where we have assumed that the time dependence is fully contained in the position and velocity variables. Of particular interest is the so-called *Action* functional $S[x(t)]$ corresponding to the time integral of the Lagrangian:

$$S[x(t)] \equiv \int_{t_0}^{t_1} L(x,v)\,. \tag{I.1.18}$$

The action is not just a function of x but a so-called functional of the function $x(t)$. You may think of the variable as being the path taken by a particle that from a position

[4] We introduce the curly or *partial derivatives* which mean that for a function of several independent variables you only take the derivative with respect to one of them (keeping the others fixed).

$x_0 = x(t_0)$ to some other position $x_1 = x(t_1)$. So if you give me $x(t)$ for all t then I can calculate $v = dx/dt$ and therefore also S, and that is why S is functional of $x(t)$. The Newtonian force law of mechanics can now be derived from a variational argument with respect to the possible paths. The variational principle says that the action is stationary under a small variation of the path. It is like saying that the extremum of a function corresponds to points where the derivative of that function is zero, indeed at a maximum or minimum of a function the slope of that function is zero. For the functional the equivalent statement is to say that an extremum for the action of a particle to go from A to B along a path corresponds to paths for which the variation in the action vanishes. So, if we make a local change of the path $x'(t) = x(t) + \delta x(t)$, then that will lead to a change in the action $S'(x(t) \equiv S(x'(t)) = S + \delta S$. The requirement that the variation $\delta S = 0$ gives rise to the so-called *Euler-Lagrange equation(s)* which reads:

$$\partial_t\left(\frac{\partial L}{\partial v}\right) - \frac{\partial L}{\partial x} = 0\,. \qquad \text{(I.1.19)}$$

One easily verifies that for the particle Lagrangian (I.1.17) one obtains Newton' second law, $m dv/dt + dV/dx = 0$. To go from the Lagrangian to the Hamiltonian formalism involves the definition of the generalized or canonical momentum p and the Hamiltonian $H(p, x)$as follows,

$$\begin{aligned} p &\equiv \frac{\partial L}{\partial v} \\ H(p,x) &\equiv pv - L\,. \end{aligned} \qquad \text{(I.1.20)}$$

There are two reasons to introduce the action one is that for relativistic systems the action is a Lorentz or relativistic invariant quantity while the energy or Hamiltonian is not, and the second has to do with quantum mechanics. There is a formulation of quantum theory, the so-called *path integral formalism* in which the quantum probability amplitude to go from x_0 to x_1 is given by a weighted sum (or integral) over all possible paths between the two points, and where the statistical weight depends on the action.

Finding the shortest path. If light or a photon goes from point A to point B, it presumably follows the shortest path and that path is a straight line between the two points. However if you use a navigator in your car, it may ask you to specify whether you mean the shortest route in a spatial sense (the cheapest) or the shortest route in time (the fastest). Knowing that 'time is money' this can be a tough decision to take.

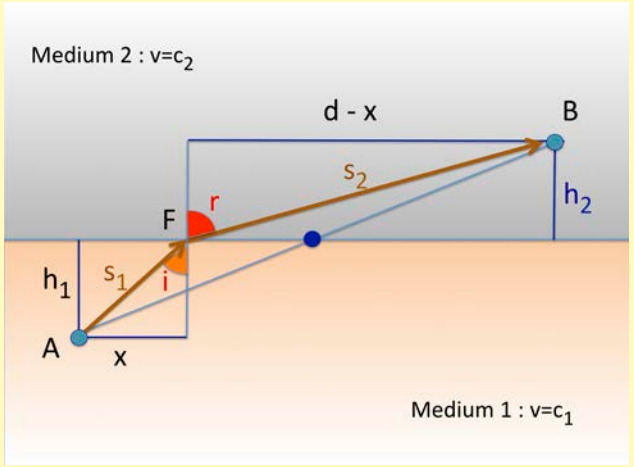
A kindergarten model for calculating the fastest path is to show kids a chopstick and stick it into a bowl filled with water. Hey! What's happening? It looks like the stick is broken! So you pull it out again and 'no' it is not broken. You hear their brains rattling. 'If there is no water it's not broken,' says one. 'It breaks at the surface,' says another. 'I think that the light ray is broken instead,' says a girl in the back. Bravo! That must be it!

The answer is given in the figure below. We have a landscape with two countries; point A is situated in the one with a maximum speed of $c_1 = 120$ km/hr and point B is in the other where the maximum speed is $c_2 = 140$ km/hr. Clearly the straight line segment AB is the spatially shortest connection. However if we want the path that takes the shortest time, we have to make a little calculation. We have indicated that the car after a distance s_1 crosses the border at a point F, which is at position x after which it goes a distance s_2 in the other country. So we choose as our action the time T it takes

to get from A to B. From the figure that

$$s_1 = \sqrt{h_1^2 + x^2} \quad \text{and} \quad s_2 = \sqrt{h_2^2 + (d-x)^2}$$

Then the calculation of T proceeds as follows:



$$T(x) = \int_A^B dt = \int_A^B (dt/ds)\, ds \qquad \text{(I.1.21)}$$

$$= \int_A^F (1/c_1) ds + \int_F^B 1/(c2)\, ds \qquad \text{(I.1.22)}$$

$$= s_1(x)/c_1 + s_2(x)/c_2\,, \qquad \text{(I.1.23)}$$

To find the minimum of $T(x)$ we have to solve for the x-value where the derivative of T vanishes:

$$\frac{dT(x)}{dx} = \frac{1}{c_1}\frac{x}{s_1} - \frac{1}{c_2}\frac{d-x}{s_2} = 0\,. \qquad \text{(I.1.24)}$$

We observe that the two quotients correspond to the sines of the angles i and r respectively, so that the condition implies the simple identity:

$$\frac{\sin i}{c_1} = \frac{\sin r}{c_2}\,, \qquad \text{(I.1.25)}$$

which is known as Snell's law for the refraction of a light ray at the interface of two media. And that brings us back to the deep connection between a broken chopstick and Google maps. After the only adult in the room had explained all this, the girl in the back still had a question: 'How does the photon know which path to choose, as I presume it doesn't know how to take a derivative?'

□

Somehow in quantum theory there are corrections to the classical picture, those are contributions that correspond to paths that are classically forbidden. ■ ■

Maxwell's electromagnetism

> It appears to me therefore, that the study of electromagnetism in all its extent has now become of the first importance as a means of promoting the progress of science.
>
> *James Clerk Maxwell, 1873*

The Maxwell equations give a unified description of electricity, magnetism and electromagnetic waves such as light or radio waves. Electromagnetism introduced the powerful concepts of a field and of field dynamics. After we discuss some of the familiar electromagnetic phenomena in relation to the Maxwell equations, we will introduce the gauge potentials which reveal two fundamental symmetries that turned out to underlie all of modern physics. The first is the so-called Lorentz invariance which lies at the root of special relativity, and the second refers to the notion of gauge invariance, a principle that underlies the description of all fundamental interactions.

Besides gravity there are basic natural phenomena of an essentially different nature to be accounted for, those re-

Figure I.1.16: *Rainbows over Holland.* Light and all its optical effects like rainbows are fully described by Maxwell's equations. Note on the right that in the barely visible secondary rainbow the sequence of colors is inverted. This is due to a third reflection of the light ray in the vapor droplets (Photo: V. de Vries).

lated to electricity and magnetism. For these the universal laws in their splendid generality were written down in a treatise by James Clark Maxwell almost two centuries after Newton's seminal contributions in about 1865. His four laws were universal as well, as they accounted for all electric and magnetic phenomena observed to that date and as a bonus turned out to also describe the propagation of electromagnetic waves in its many guises such as light, radio waves or X-rays. Maxwell created for us the grand synthesis of many of the laws that were proposed earlier on by Coulomb, Ampère, Faraday, Lenz and many others. And a unified picture emerged of what once were considered entirely disconnected phenomena: electricity, magnetism and optics.

Electromagnetic Fields. Maxwell's theory is formulated in terms of a magnetic field $\mathbf{B}$ and electric field $\mathbf{E}$, which depend on space and time. So at any instant in time at any point in space, the fields have a particular strength $(\mathbf{E}(x,t),\mathbf{B}(x,t))$. You may think of them as two little arrows (vectors) pointing in some directions in space. The Maxwell laws describe in detail how electric currents cause magnetic fields, and how changes in magnetic flux result in currents which counteract that change. The laws also describe how accelerated charges emit electromagnetic radiation. From a more formal point of view they brought the fundamental but rather abstract concept of a *field* to life, in the sense that this concept was promoted from a mere mathematical abstraction and calculational tool to a physical reality. Electromagnetic fields by themselves propa-

gate through space and time as waves and radiation, and turned into physical entities carrying energy and momentum. When you spend a day on the beach and forgot your sunscreen, you learn the hard way how much energy the electromagnetic waves emitted from the sun can carry. But also the beauty of a rainbow on a both sunny and foggy day is a manifestation of the electromagnetic interaction of light rays with the tiny vapor droplets in the fog.

The Maxwell equations

I am going to write down the Maxwell equations in their full glory: in other words, in their gory detail. Not to scare or impress you but because they are truly iconic. I think you need to have seen them, otherwise it is like going to Paris for the first time and missing out on the Eiffel tower, that would presumably make you mad at your tour operator. The comments I will make are rather general and descriptive which hopefully makes showing them less daunting. These are the four equations that could equally well be called 'the four Maxwell laws of electromagnetism and light.' These equations are usually presented in the following form:[5]

$$\boldsymbol{\nabla}\cdot\mathbf{E}=\rho\,, \tag{I.1.26a}$$

$$\boldsymbol{\nabla}\cdot\mathbf{B}=0\,, \tag{I.1.26b}$$

$$\boldsymbol{\nabla}\times\mathbf{B}=\frac{\mathbf{j}}{c}+\frac{1}{c}\frac{\partial\mathbf{E}}{\partial t}\,, \tag{I.1.26c}$$

$$\boldsymbol{\nabla}\times\mathbf{E}=-\frac{1}{c}\frac{\partial\mathbf{B}}{\partial t}\,. \tag{I.1.26d}$$

We see that the equations, besides the $\mathbf{E}$ and $\mathbf{B}$ fields, depend on the charge and current densities $\rho(x,t)$ and $\mathbf{j}(x,t)$, and on the velocity of light c. That the charges and currents appear in these equations is no surprise as they are the *sources* of the fields.

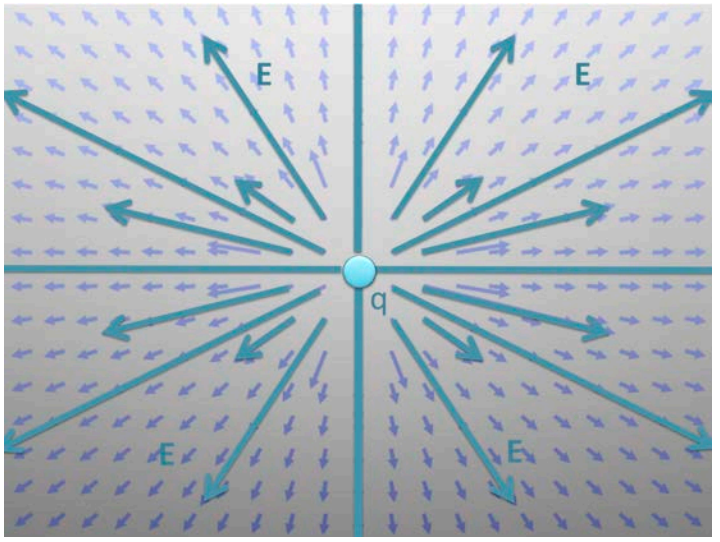


Figure I.1.17: *Coulomb's law.* If we put a positive charge at rest at the origin, then the first Maxwell equation corresponding to *Coulomb's law* will yield an electric field pointing radially outward. The strength of the field (given by the length of the vector) falls off as $1/r^2$ in three dimensions. This equation by itself describes what is called *electro-statics*. The second Maxwell equation tells you that the magnetic equivalent of such a radial field does not exist.

The *first equation* is often called *Coulomb's* or *Gauss' law*, and it determines the electric field that is caused by a given charge distribution. It says in particular that a single charge causes a radial electric field around it, as illustrated in Figure I.1.17.

The *second equation* is the magnetic analogue of the first equation for isolated magnetic 'North' or 'South' charges. The right-hand side is put to zero, for the excellent reason that magnetic monopoles have never been observed, at least up until now. This is the *'no monopole' equation*, but one sees that the system could be adapted to a situation where monopoles would show up, a situation that cannot be excluded *a priori*.

The *third equation*, also called *Ampère's law*, states that a current (or moving charge) causes magnetic fields and

[5] The way they look depends on the precise choice of units, here I work in Heaviside-Lorentz units because that choice makes them look simpler. The physical parameter is the velocity of light c, and as we will see it will pop up in most relations.

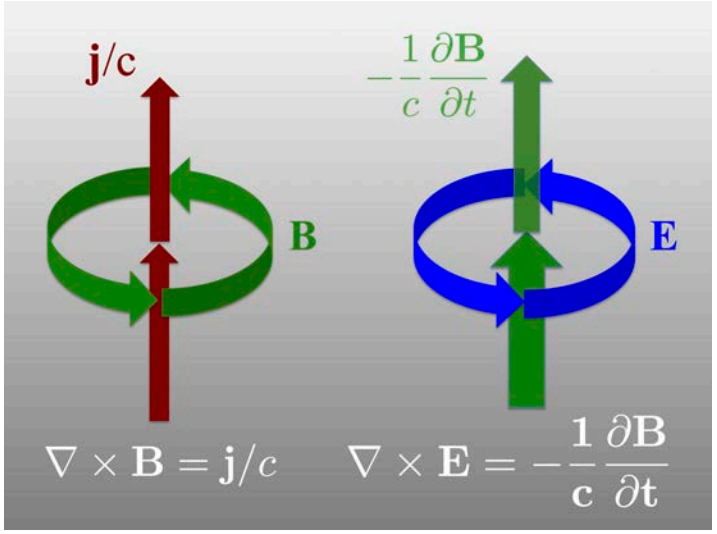


Figure I.1.18: *Ampère's and Faraday's laws.* This figure illustrates the two Maxwell equations involving the curl of the fields. The left picture refers to *Ampère's law* for the case of magnetostatics, where a straight current yields an axially symmetric B field. The picture on the right depicts *Faraday's* or *Lenz's law* describing how a changing magnetic field or flux gives rise to an electric field. If we think of the $\mathbf{E}$ loop as a closed conducting loop, a current would start flowing so as to counteract the change in the magnetic field.

a changing electric field. In other words, given the distribution of charges and currents in space and time, the Maxwell equations tell you exactly what the electromagnetic fields will look like. The third and fourth equation involve the so-called curl of a magnetic and electric field. In Figure I.1.18 we have indicated how the fields indeed 'curl' around the source which is a vector like the current. It is another instance of the 'Like-rule.'

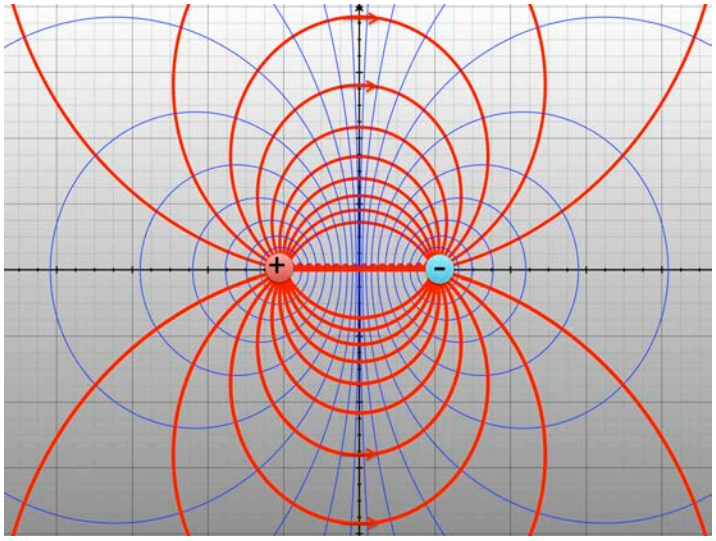
The *fourth equation*, also called *Faraday's* or *Lenz's law*, describes how a changing magnetic field causes (induces) an electric field, which in turn can give rise to a current. If you take a conducting loop and you change the magnetic flux through that loop, then that change induces a current through the loop. If the loop is made of a superconducting material, the current will keep running forever. Also in this equation we note that a potential 'magnetic current term' is manifestly absent for the same reason as before. It is this absence of magnetic monopoles and currents that breaks the would-be symmetry between electric and magnetic phenomena.

All the magnetic phenomena we have observed up to now are understood as caused by currents, meaning moving electric charges. Indeed, the second equation tells you that there are no magnetic purely radial monopole fields, while the third equation tells you that if you make a tiny closed current loop, it will act like a tiny magnetic dipole, and the overall configuration is a magnetic 'dipolar' field. You guessed it: all real magnets correspond to zillions of microscopic current loops, all neatly lined up. With the well-known consequence that if you break a bar magnet in half, you do not get a separated North and South pole, you just get two smaller dipolar bar magnets.

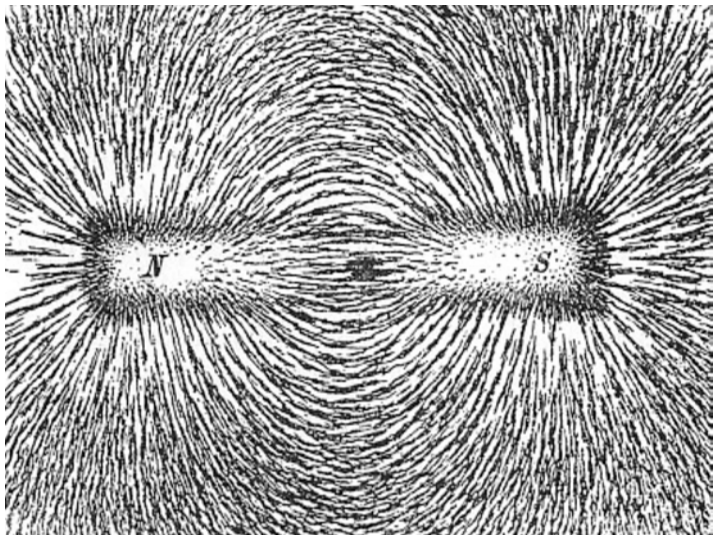
Linearity. It is important to observe that the system of Maxwell equations is linear in the fields. This means that one can simply add different solutions. In other words if I have any set of solutions, then any linear combination of these would again be a solution. This is illustrated in Figure I.1.19 This linearity of the dynamical system basically means that the electromagnetic field does not interact with itself.

Electric-magnetic duality. We have emphasized that the asymmetry of the Maxwell equations reflects the asymmetry of nature with respect to the existence of electric charges and magnetic monopoles. Indeed if we restrict the equations to a source-free situation, meaning that ρ and $\mathbf{j}$ are zero, then the equations exhibit a manifest symmetry which is referred to as *electric-magnetic duality*. The system of equations is in that case invariant under the dual transformation or mapping, where we simultaneously make the replacements $\mathbf{E} \to \mathbf{B}$ and $\mathbf{B} \to -\mathbf{E}$. This mapping transforms the first pair of equations into each other, and similarly likewise the second pair.

Light as an electromagnetic wave. The most impressive



(a) Electric dipole field resulting from two opposite charges. The electric field lines are red, and the equipotential lines are blue.



(b) Magnetic dipole field caused by a bar magnet. They are made visible byputting the magnet on a table and spread some iron filings around it.

Figure I.1.19: *Dipolar fields.* If we put two opposite point charges at some distance of each other, the resulting field becomes dipolar, meaning that the field lines start at the positive charge (magnetic north) and end on the negative charges. In (a) we have the electric dipole field and in (b) we have the magnetic example. The second is approximated by an ordinary dipolar bar magnet. The field configuration is because of the linearity obtained by just adding at every point the two coulomb fields of the single charges as depicted in Figure I.1.17

and surprising achievement of Maxwell was the great discovery that even in the absence of sources, the equations allowed for solutions describing electromagnetic waves that propagate through empty space at the velocity of light. This explains why the only parameter that appears in these equations is the velocity of light. We will return to these electromagnetic waves shortly.

It is gratifying to see how much 'truth' about physical reality can be described with so few symbols. You could say that the ultimate elegance of nature is most manifest once it is expressed in the powerful language of mathematics. Awesome indeed!

Partial differential equations. The equations form a system of partial differential equations, partial because the fields depend on space and time variables, and the derivatives that appear are with respect to the spatial coordinates as well as time. This explains also the appearance of the *'del'* or *'nabla'* operator $\boldsymbol{\nabla}$, which is just the 'vector of spatial derivatives,'

$$\boldsymbol{\nabla} = \{\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \frac{\partial}{\partial x_3}\}. \qquad \text{(I.1.27)}$$

To systematically solve equations involving the vector operator $\boldsymbol{\nabla}$, mathematicians have developed a special subject called *vector calculus*. That is what physics students have to study and are supposed to master, and as such, it is far beyond the scope of this book. You will believe me if I say that many shelves in our university libraries are full of books and journals that are stuffed with explicit solutions of the Maxwell equations for virtually any imaginable situation. With all due respect, we will stay far from those impressive halls of wisdom, though we discuss some funda-

mental theorems involving the nabla operator $\boldsymbol{\nabla}$ in a *Math Excursion* at the end of Part III on page 621. My narrative only tries the convey the overall structural aspects of the theory, which by the way does not force my story to become superficial, in fact quite the contrary.

A dynamical systems perspective. We may elevate the dynamical systems' pespective of the previous section on mechanics to the Maxwell equations and say that the dynamical 'variables' are now the components of the $\mathbf{E}$ and $\mathbf{B}$ fields which satisfy certain dynamical equations or equations of motion,

$$\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}t} = \mathrm{f}_{\mathrm{B}}(\mathbf{E},\mathbf{B})\,, \tag{I.1.28a}$$

$$\frac{\mathrm{d}\mathbf{E}}{\mathrm{d}t} = \mathrm{f}_{\mathrm{E}}(\mathbf{E},\mathbf{B})\,. \tag{I.1.28b}$$

Locality. These are indeed only two of the four Maxwell equations, those with time derivatives in them. Note that on the right-hand side I have for convenience suppressed the dependence on the spatial derivatives of the fields, because at a given time t these can be calculated from the field themselves at time t.The main point here is that the equations are local: loosely speaking one could consider the fields as an infinite collection of independent variables which are only locally coupled.

Constraints. The other pair of equations without time derivatives are constraint equations; in order for the system to be consistent, these have to be obeyed at all times. So if these equations are satisfied at some initial time $\mathrm{t} = 0$, then consistency of the system requires that they remain valid for all t, and this requires that the time derivatives of those equations should vanish.

This, in turn, can be proven from the Maxwell equations. For the second equation the argument is quite straightforward: one finds that by taking the time derivative of that equation one obtains the same expression as by taking the divergence of the right-hand side of the fourth equation. The latter, in turn, equals $\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\mathbf{E})$, which vanishes identically, meaning that it is zero for any field $\mathbf{E}$. This is discussed in the *Math Excursion* on vector calculus on page 621 of Part III. For the first and third Maxwell equations a similar argument can be applied, comparing the time derivative of the first and the divergence of the third equation we see that consistency requires the following relation to hold:

$$\frac{\partial\rho}{\partial t} = \boldsymbol{\nabla}\cdot\mathbf{j}\,. \tag{I.1.29}$$

This equation is the continuity equation for electric charge, it relates the time derivative of the charge in a given volume with the current through the surface bounding that volume. In other words, it is the local conservation law for electric charge. The conclusion is that the consistency of the Maxwell equations requires local charge conservation.

Constraint equations can be used to reduce the number of independent degrees of freedom, fields in this case. What that means is that electromagnetism does not really have two times three equals six independent field components as the two equations above suggest. Maxwell's first and second equations express two local – (x,t) dependent 'constraints,' which reduce the number of independent field variables from six to four. And these correspond to the four gauge potentials we will get to shortly. Nevertheless, from this dynamical systems point of view there is a remarkable structural similarity between the mechanical and electromagnetic systems.

The electromagnetic force exerted on a charge. The Maxwell equations feature external sources in terms of charges and currents. Clearly these refer to charged particles or collectives thereof. So to complete the dynamical system approach we should also include the dynamics of the charges and currents. This in turn means that we specify the forces that these are subject to in given electric and magnetic fields. The expression for this so-called

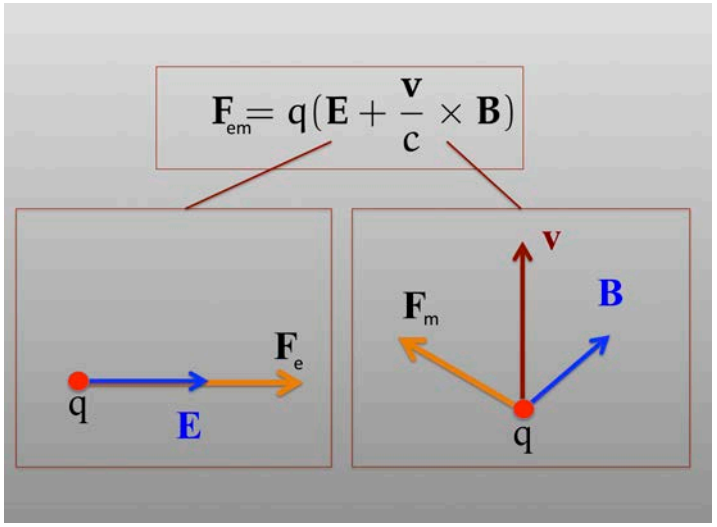


Figure I.1.20: *Motion of charge in an electromagnetic field.* This figure illustrates how the Lorentz force works on a charged particle. We show that the force has two contributions: one proportional to and in the direction of the electric field and one proportional to the magnetic field and the velocity in a direction perpendicular to the field and the velocity.

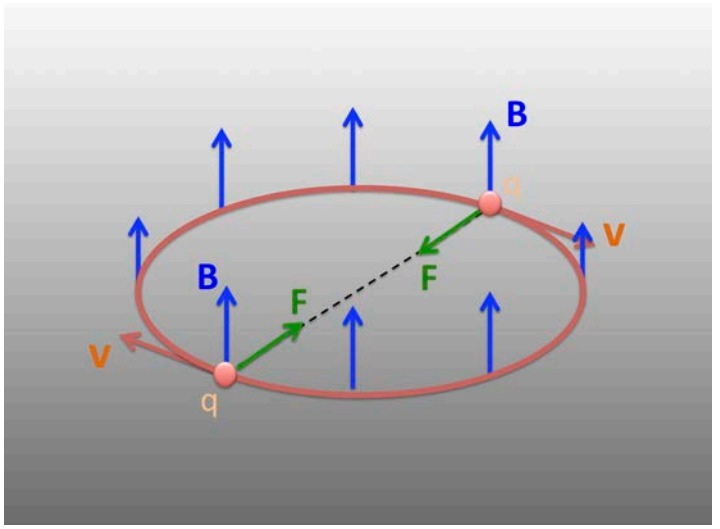


Figure I.1.21: *Motion of charge in a constant magnetic field.* This figure shows the orbit of a charged particle with a velocity perpendicular to the field. The force is constant and perpendicular to the velocity and will cause the particle to have a circular orbit. As the force is always perpendicular to the orbit the magnetic field does not do any work, and the magnitude of the velocity remains constant.

Lorentz force exerted on a charge at a point $(\mathbf{x}, t)$ by the fields $\mathbf{E}(\mathbf{x}, t)$ and $\mathbf{B}(\mathbf{x}, t)$ is the following:

$$\mathbf{F} = q(\mathbf{E} + \frac{\mathbf{v}}{c} \times \mathbf{B})\,. \tag{I.1.30}$$

The first term is a force in the direction of the electric field that any charge will feel, and the second term is the magnetic, so-called *Lorentz force*, which is orthogonal to the velocity of the charged particle. It is proportional to the magnitude of the current $\mathbf{j} = q\mathbf{v}$ and clearly vanishes when a particle is at rest. The fact that the magnetic component of the force is perpendicular to the velocity means that that component is always perpendicular to the trajectory, and consequently implies that the magnetic field cannot do any work on the charge. A charge in a constant magnetic field perpendicular to its velocity would therefore move in a circular orbit as we depicted in Figure I.1.21.

Clearly, the dynamical system to be solved is the coupled system of Newton's and Maxwell's equations where Newton's equations have to include the Lorentz force and the charge(s) and their currents have to be included as sources in the Maxwell equations. This system is of course non-linear because of the feedback caused by the interaction terms.

We will later show how the electromagnetic interaction affects the energy function or the Hamiltonian of a charged particle, but that is more conveniently expressed in terms of the gauge potentials that we will introduce shortly.

Field energy and momentum. If we put a charged particle in a constant electric field, the field will exert a constant force on the particle which will therefore start to accelerate uniformly. This in turn means that its energy will increase. Now if we want to maintain the sacred principle of overall energy conservation, then one is forced to assume that the electromagnetic field also carries energy.

Indeed, the mere fact that the Maxwell equations without any charges and currents describe propagating waves means that the fields should carry both energy and momentum. Furthermore, once properly defined, it turns out that both the total energy and momentum for the whole system including charges and currents and fields is conserved again, assuming of course that the fields evolve according to Maxwell's equations.

Because the electric and magnetic fields as fundamental variables are space-time dependent – we say that they describe *local* degrees of freedom, it is then natural to define field energy and field momentum *densities*. This means that in order to get the total energy/momentum within a given volume one has to integrate the densities over that volume.
The expression for the energy of the electromagnetic field is basically the sum of (or better the integral over) the contributions in all points in space of a field energy density

$$\varepsilon(x,t) \;=\; \frac{1}{2}(|\mathbf{E}|^2 \;+\; |\mathbf{B}|^2)\,,$$

which is quadratic in $\mathbf{E}$ and in $\mathbf{B}$, where you may think of the first term as corresponding to the 'kinetic energy' and the second to the 'potential energy.' This total energy is conserved. The fields also carry a momentum density, which is called the *Poynting vector* $\mathbf{S}(x,t) \;=\; c\,(\mathbf{E} \;\times\; \mathbf{B})$ and an angular momentum density $\mathbf{L}(x) \;=\; (\mathbf{x} \;\times\; \mathbf{S})/c^2$ in complete analogy with particle angular momentum $\mathbf{L} = \mathbf{x} \;\times\; \mathbf{p}$. This comes out most clearly in the electromagnetic wave solutions to the Maxwell equations illustrated in Figure I.1.23, which shows that the fields form propagating waves that are transversal, meaning that at any point in space the vectors $\mathbf{E}$ and $\mathbf{B}$ are mutually perpendicular, and also perpendicular to the direction of propagation. From the figure one verifies that the field momentum density $\mathbf{S}$ is, as expected, directed along the propagation direction of the wave.

Three fundamental principles. The remainder of this section is devoted to two fundamental symmetry principles underlying the Maxwell equations of electromagnetism.
The *first principle* refers to the notion of *Lorentz invariance* which forms a key link with the theory of relativity.
The *second principle* refers to the notion of *gauge invariance* which amounts to a hidden redundancy that is present if we describe electromagnetism in terms of $\mathbf{E}$ and $\mathbf{B}$ fields as we usually do.
The *third principle* concerns the quantum nature of electromagnetism, of which the most basic manifestation is that we have to think of electromagnetic fields in terms of particle-like excitations or quanta, called photons. The latter principle is the main subject of the book and will be fully explored in the forthcoming chapters; we will not discuss it any further here.

The Maxwell equations refer to the fields $\mathbf{E}$ and $\mathbf{B}$, because these fields are the physical fields we can measure quite directly. The equations are beautiful, but that beauty has its price in the sense that the description is highly redundant and therefore basically inefficient! The reason we already touch on these rather sophisticated symmetry principles here is that in hindsight it turns out that these two invariances, combined with the principles of *quantum theory*, really form the conceptual backbone of all of modern fundamental physics. The tremendously successful Standard Model of fundamental forces and particles is a particular expression of these three underlying principles. Moreover, understanding these principles played an essential guiding role in discovering the Standard Model.

Electromagnetic waves

The source-free Maxwell equations can be recast in the form of wave equations. The wave equations manifestly display the underlying Lorentz or relativistic invariance of the Maxwell theory. In that sense Maxwell theory was the cradle of relativity.

Relativistic wave equations. 
By mathematically manipulating them we can cast the Maxwell equations (I.1.28) in an alternative form. In the case of vanishing sources – with zero charges and currents in other words – they take the form of two wave equations: one for the electric and and one for the magnetic field.[6]

These wave equations are Lorentz and therefore relativistically invariant, which means, as we will discuss later in the corresponding section on page 60, that they will take the same form for different observers that move at a constant speed with respect to one another. Such observers have coordinate frames that are different, but the statement is that the frames of two such observers are related by a so-called Lorentz transformation, which depends on their relative velocity. An alternative way to express the fact that the equations 'look the same' for the different observers is to say that the equations are invariant under Lorentz transformations.

Four-vectors. Let us look at this a little closer. In ordinary space we can define a coordinate vector $\mathbf{x}$, and then we know that a rotation will change the direction it is pointing. What does not change is the dot product or the length of the vector, $\mathbf{x}\cdot\mathbf{x} = x^2$. The length of any vector is invariant under rotations, and this also holds therefore for the square of the vector operator $\boldsymbol{\nabla}$. To explain the notions of Lorentz invariance and of space-time we do something similar. First we define a space-time coordinate *four-vector* $x^\mu = \{x^0, \mathbf{x}\}$ with $x^0 \equiv ct$, the factor c is there to also give x_0 the dimension of a length. Next we define the relativistic 'length' or *space-time interval* s of that coordinate vector by the relation $s^2 \equiv x^\mu x_\mu \equiv x_0^2 - \mathbf{x}\cdot\mathbf{x}$, where indeed the repeated upper and lower μ index by definition means that we have to sum over its range $0, ..., 3$, with the minus sign for the spatial components included. The notion of Lorentz invariance refers now to the fact that the space-time interval is invariant under Lorentz transformations, just like the length of an ordinary vector is invariant under rotations. So Lorentz transformations are the generalization of ordinary rotations in three-dimensional Euclidean space to four-dimensional space-time (also called Minkowski space).

[6] A typical 'wave equation' is discussed in the *Math Excursion* at the end of Volume III on page 613.

Figure I.1.22: *Aurora Borealis.* The *Northern Lights* are caused by collisions of charged particles coming from the sun and gas particles from the earth's atmosphere. The most common auroral color, a pale yellowish-green, is produced by oxygen molecules located about 60 miles above the earth. Rare, all-red auroras are produced by high-altitude oxygen, at heights of up to 200 miles. (Source: Wikimedia)

The box-operator. The wave equations feature second order spatial and time derivatives in a unique relativistically invariant combination denoted by

$$\Box \equiv \partial^\mu \partial_\mu \equiv \frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \boldsymbol{\nabla}^2 . \qquad \text{(I.1.31)}$$

The electromagnetic wave equations can then simply be written as

$$\Box \mathbf{E} = 0 , \qquad \text{(I.1.32a)}$$

$$\Box \mathbf{B} = 0 . \qquad \text{(I.1.32b)}$$

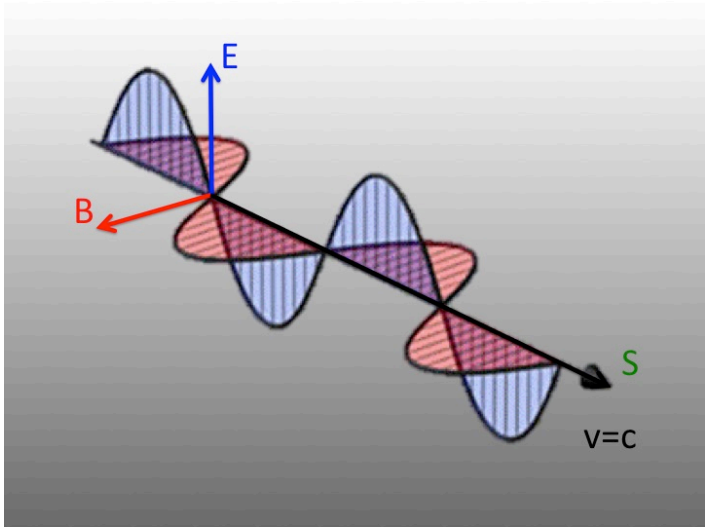


Figure I.1.23: *Electromagnetic wave.* This is a propagating wave of periodic electric and magnetic fields. The polarizations of the electric and magnetic field are orthogonal, and both are orthogonal to the direction of propagation which is along the direction of the field momentum $\mathbf{S}$.

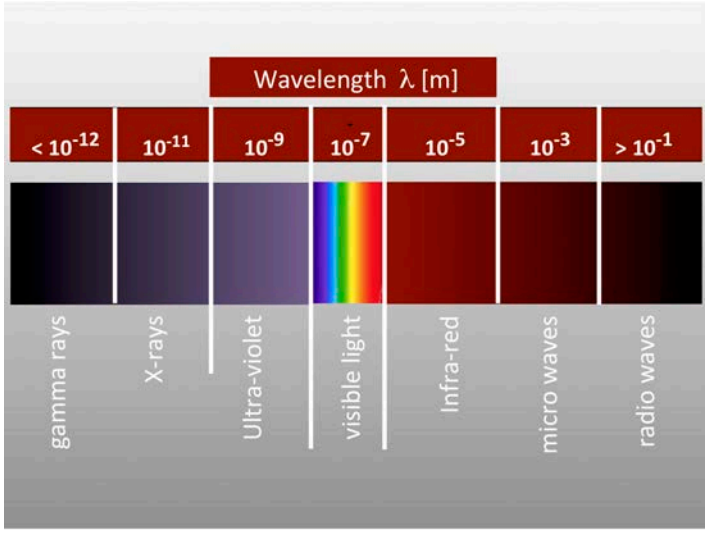


Figure I.1.24: *Electromagnetic radiation spectrum.* Classical electromagnetic waves can have any wavelength, from very long wavelength radio waves to the ultra short wavelength hard gamma rays. Visible light represents a narrow range in the center.

In the 'box operator' $\Box$ we see that time and space appear on an equal footing, which amounts to saying that this operator is relativistically invariant. The 'box' operator is the relativistic wave operator, and the equations above are the equations for electromagnetic waves. And indeed, it was this property of invariance of the Maxwell equations under the Lorentz transformations, named after its discoverer, the Dutch physicist and early Nobel laureate Hendrik Antoon Lorentz, which was a crucial key used by Einstein to unlock the gateway to the world of relativity.

Basic properties of waves. Like all waves, the electromagnetic waves are characterized by a wavelength λ, a frequency ν, and a velocity $\mathbf{v}$ which in this case of course equals the speed of light, $|\mathbf{v}| = c$. These three quantities are not independent, since they satisfy the relation $\nu = c/\lambda$. So electromagnetic waves are special in that they always travel with the speed of light, you can't speed them up or slow them down. If you put more energy into the waves, two things may happen: (i) the amplitudes of components may go up (the signal becomes more intense), and/or (ii) the frequency may increase, meaning that the colour (in the case of light) will be shifted towards the blue. In the quantum world where we think of photons or particles of light, the corresponding mechanisms are, (i) that we can create more particles of light, or (ii) we can give the particles themselves more energy by increasing the frequency.

We have depicted the characteristic spatial structure of a classical electromagnetic wave in Figure I.1.23, and one sees that for such waves the directions (or polarizations) of the electric and magnetic field amplitudes are orthogonal and orthogonal to the propagation direction as well. The discovery of these wavelike solutions was a seminal contribution to electromagnetic theory, because it unified electromagnetism with the field of optics. The waves can in principle have any frequency or wavelength. We have

sketched the spectrum of electromagnetic radiation in Figure I.1.24 , from which we see that spectrum of visible light only covers a narrow range in the center. At the long wavelength side the spectrum continues via the infrared into the micro and radio waves. On the short wavelength side it continues in the ultraviolet via X-rays into hard gamma rays. This side of the spectrum corresponds to ionizing radiation, where ionizing means that the electrons in the outer shells of atoms and molecules will be kicked out so that positively charged ions stay behind. This among other things means that this radiation is very damaging to biological tissue and one should avoid being exposed to it. In other words, avoid spending the weekend on a tropical beach without sunscreen.■

Lorentz invariance: the key to relativity 

We introduce the electromagnetic gauge potentials, and rewriting the electromagnetic fields in terms of these reduces the number of independent equations to four. In this form the invariance of the system under Lorentz transformations becomes manifest, establishing that the system is fully relativistic. This and the following section basically show a form in which the Maxwell equations can be cast that maximally exhibits their fundamental structure and beauty.

Gauge potentials. It is interesting that in the context of quantum theory it is far more profitable to use a different parametrization of the electromagnetic field in terms of so-called *gauge potentials* denoted by $A_\mu(x,t)$. As before, the index μ runs from $0,...,3$, with 0 the time component and $1,2,3$ the space components.

The four-vector $A_\mu = (V, -\mathbf{A})$ are the electromagnetic potentials where V is often referred to as the electrostatic or *scalar potential* and $\mathbf{A}$ as the *vector potential*. From these potentials the electric and magnetic field can be calculated directly through the defining relationships:

$$\mathbf{B} = \boldsymbol{\nabla} \times \mathbf{A}\,, \tag{I.1.33a}$$

$$\mathbf{E} = -\boldsymbol{\nabla} V - \frac{1}{c}\frac{\partial}{\partial t}\mathbf{A}\,. \tag{I.1.33b}$$

Let us indicate how these expressions come about. One may show that for any magnetic field configuration $\mathbf{B}$ with zero divergence, meaning that it satisfies equation (I.1.26b), there is a vector field $\mathbf{A}$ that satisfies equation (I.1.33a). In fact that $\mathbf{A}$ is not unique as we'll see later. Indeed one finds that the equality $\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \mathbf{A}) \equiv 0$ holds for *any* $\mathbf{A}$; it is a mathematical identity which basically follows from the definition of the vector derivative $\boldsymbol{\nabla}$. If we proceed by substituting this expression of $\mathbf{B}$ into the equation (I.1.26d), we get an equation of the type $\boldsymbol{\nabla} \times \mathbf{C} = 0$, with $\mathbf{C} = \mathbf{E} + \frac{1}{c}\frac{\partial}{\partial t}\mathbf{A}$. Now there is another identity that says that any field $\mathbf{C}$, whose rotation vanishes, can be written as a gradient of some scalar field V. This means that we may write $\mathbf{C} = \boldsymbol{\nabla} V$, from which the equation (I.1.33b) then follows. So by changing from the $\mathbf{E}$ and $\mathbf{B}$ fields to the potential $A_\mu = (V, -\mathbf{A})$ we have identically satisfied two of the four Maxwell equations. From the other two follow equations that the gauge potentials have to satisfy.

The electromagnetic field strength. You might wonder why I – clearly being in love with relativity – don't come up with four vectors E_μ and B_μ . Alas, 'It ain't necessarily so....' Better even, 'it just ain't gonna work!' The appropriate relativistic place for the electric and magnetic fields is that they correspond to the components of an antisymmetric two index object (a tensor) called the *field strength* $F_{\mu\nu}$:

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu\,. \tag{I.1.34}$$

The three spatial components F_{ij} correspond with the components of $\mathbf{B}$, and the space-time components F_{0i} correspond with the components of $\mathbf{E}$. The $\mu - \nu$ antisymmetry can be visualized more conveniently by writing F as an an-

tisymmetric 4×4 matrix:

$$F = \begin{vmatrix} 0 & E_1 & E_2 & E_3 \\ -E_1 & 0 & -B_3 & B_2 \\ -E_2 & B_3 & 0 & -B_1 \\ -E_3 & -B_2 & B_1 & 0 \end{vmatrix} . \qquad \text{(I.1.35)}$$

It clearly shows how the components of the $\mathbf{E}$ and $\mathbf{B}$ fields are not part of four vectors, which means that the $\mathbf{E}$ and $\mathbf{B}$ components may mix if we make a Lorentz transformation from one reference frame to another, just like the space and time components of the position four-vector do. This mixing is not entirely unexpected, since if we can transform a particle at rest in one frame to a moving particle in another frame, then the static particle has a pure radial electric field. The moving charge, however, is like a current and generates a magnetic field as well. So one expects that under a Lorentz transformation the $\mathbf{E}$ and $\mathbf{B}$ fields should mix. And if each of them was a four-vector, transformations would *not* mix the two sets of components.

From the manifestly relativistic definitions above, we see that the symmetry between electric and magnetic fields is particularly special to four-dimensional space-time. If we consider what the matrix $F_{\mu\nu}$ would look like in different dimensions, this becomes very clear: (i) in two-dimensional space-time there is only a single electric field component along the space direction and there is no magnetic field; (ii) in three dimensions we have an electric vector field with two components and a single component magnetic field which is therefore like a (pseudo) scalar.

We can now also write the Maxwell equations in manifestly relativistic form. The equations with sources (I.1.26a) and (I.1.26c) will then read:

$$\partial^\nu F_{\mu\nu} = \frac{1}{c} j_\mu \, , \qquad \text{(I.1.36)}$$

where a repeated upper and lower index implies a summation over that index from $0, ..., 3$. On the right-hand side we have the current j_μ, which is now also a four-vector. Its time component j_0 is equal to the charge density ρ times the velocity of light c, and the spatial components j_i are the components of the usual electric current-density vector $\mathbf{j}$:

$$j_\mu = (c\rho, \mathbf{j}). \qquad \text{(I.1.37)}$$

The other two – sourceless – Maxwell equations can also be written in a manifestly Lorentz invariant way as,

$$\partial^\nu \widetilde{F}_{\mu\nu} = 0 \, . \qquad \text{(I.1.38)}$$

Where we have constructed the *dual field strength* $\widetilde{F}_{\mu\nu}$ marked with a 'tilde,' by applying the electric-magnetic duality transformation discussed on page 22, to $F_{\mu\nu}$, yielding,

$$\widetilde{F} = \begin{vmatrix} 0 & B_1 & B_2 & B_3 \\ -B_1 & 0 & E_3 & -E_2 \\ -B_2 & -E_3 & 0 & E_1 \\ -B_3 & E_2 & -E_1 & 0 \end{vmatrix} . \qquad \text{(I.1.39)}$$

Again these sourceless equations are solved identically by substituting the field strength in terms of the gauge potentials. In other words, by substituting the expressions (I.1.33) of $\mathbf{E}$ and $\mathbf{B}$ in terms of the gauge potentials into the equation (I.1.38).

The action for the Maxwell field. We have, in the closing subsection about classical mechanics, highlighted the importance of the concept of an action (and Lagrangian) for relativistic systems. As the Maxwell system is a relativistic system with the fields and their derivatives as fundamental degrees of freedom, we should ask whether there is a suitable form of the Lagrange formalism in this case. The answer is affirmative, so let us show what it looks like. First of all let us introduce the Lagrangian density which corresponds to the Lorentz invariant expression that is quadratic in the derivatives of the field:

$$\mathcal{L}(A_\mu, \partial_\nu A_\mu) = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} - j_\mu A^\mu \, . \qquad \text{(I.1.40)}$$

The Lagrangian L would be given by the integration over space of the density $\mathcal{L}$, and the action S is obtained by

an additional integration over time. This yields the fully covariant expression,

$$S[A_\mu] = \int \mathcal{L}(A_\mu, \partial_\nu A_\mu)\, d^4x\,. \qquad \text{(I.1.41)}$$

One may show that the Maxwell equations (I.1.36) correspond to the Euler-Lagrange equations for this action.

Current conservation. The previous equations require that the current j_μ is conserved, which means to say that

$$\partial^\mu j_\mu = 0\,. \qquad \text{(I.1.42)}$$

The substitution of the definitions yields the continuity equation which expresses the local conservation law for electric charge,

$$\frac{\partial \rho}{\partial t} = \boldsymbol{\nabla}\cdot\mathbf{j}\,. \qquad \text{(I.1.43)}$$

Integrating this equation over some volume $\mathcal{V}$, it states that the increase of the charge in $\mathcal{V}$ per unit time (and divided by c) equals the net electric current flowing inward through the closed surface that bounds that volume.

A way to think about this is to consider an office building where people go in and out. Then if we state that the number of people in the building is locally conserved, it means that the total number of people in the building is equal to the number that are already in there, plus or minus the people who enter or leave the building. It is local because you can apply it to any volume, for example the law also applies to any floor of the building, or any individual room for that matter.

The energy of a charged particle. A good reason to introduce the gauge potentials is that the coupling of the electromagnetic field to charged particles and fields takes a particularly simple form. The correct expression for the interaction with a charged particle is directly obtained by replacing, in the non-interacting particle theory, the momentum vector $\mathbf{p}$ of the particle by $\mathbf{p} + q\mathbf{A}/c$, and the energy E by $E - qV$, where q is the charge of the particle. The energy function or Hamiltonian H for the charged particle simply becomes:

$$H - qV = \frac{1}{2m}(\mathbf{p} + \frac{q}{c}\mathbf{A})^2\,. \qquad \text{(I.1.44)}$$

From this expression for the Hamiltonian, one obtains the equation of motion for a charged particle, which yields as one might expect the Newton force law featuring the Lorentz force:

$$\frac{d\mathbf{p}}{dt} = q(\mathbf{E} + \frac{1}{c}\mathbf{v}\times\mathbf{B})\,. \qquad \text{(I.1.45)}$$

What has become clear from my exposition so far is that the electromagnetic 'field' as we know it in classical physics basically corresponds to a system with an 'infinite' number of degrees of freedom, namely the $\mathbf{A}$, or $\mathbf{B}$ and $\mathbf{E}$ fields that can vary at any point in space, so that a field represents a degree of freedom in any point of space. We have emphasized the dynamical systems perspective because it is significant if we consider the quantum theories of fields and want to compare them to the quantum theory of particles. ■

The charge degree of freedom. If we speak of 'a charge,' we commonly imagine a point-like particle carrying a certain charge, and as far as we know that charge q is quantized in units of the fundamental electron-charge $-e$. If the charge q has a velocity, it corresponds to a current $\mathbf{j} = q\mathbf{v}$, localized at the position of the particle. Often, though, we think of a *charge density* which is taken to be a continuous distribution.

The charge and current density $(c\rho, \mathbf{j})$ become the charge and current of a point charge q and $\mathbf{j}$, multiplied by a distribution function f^2, which specifies how the charge and currents are spread around $x_\mu(t)$.

A preliminary leap into quantum mechanics. At this point it may be illuminating to jump ahead into the quantum domain where things are so very different. For one thing, in quantum theory a charged particle is represented by a

the complex function, the so-called wavefunction $\Psi(\mathbf{x},t)$, which describes the quantum states of the particle. 'Complex' here means that the wavefunction has a 'real' and 'imaginary' part, and we may write the wavefunction therefore as $\Psi(\mathbf{x},t) = e^{-i\alpha}\, f(\mathbf{x},t)$, the product of a local factor with a phase $\alpha(\mathbf{x},t)$ and a real function $f(\mathbf{x},t)$. Whereas the state at some time t of a classical particle is determined by specifying its position, velocity (and parameters like mass and charge), in quantum physics the state is specified by the wavefunction which is defined over all of space. This means that the Lorentz equation of motion for a charged particle (I.1.45) will turn into the famous Schrödinger equation for the wavefunction Ψ. Quite a difference indeed, and in the second Volume of the book we will fully explore what it all implies.

In quantum theory for a single particle the momentum $\mathbf{p}$ is represented by the differential operator $\mathbf{p} = -i\hbar\boldsymbol{\nabla}$ which is supposed to act on the wavefunction. And it is basically here that the famous Planck constant 'h-bar' $\hbar \equiv h/2\pi$ enters the mathematical formalism. The coupling with the vector potential is, as mentioned before, implemented following the minimal replacement $\mathbf{p} \;\Rightarrow\; \mathbf{p} + q\mathbf{A}/c$, meaning the in the quantum world we have to replace the ordinary vector derivative $\boldsymbol{\nabla}$ with the *covariant derivative* $\mathbf{D} \equiv \boldsymbol{\nabla} + iq\mathbf{A}/\hbar c$.

The distribution $|\Psi|^2 = \Psi^*\Psi = f(\mathbf{x},t)^2$ represents the *probability density* of finding the particle at the position $\mathbf{x}$ upon a position measurement at time t. This distribution is independent of the phase α. So it is not the charge which is distributed over space, it is the probability of finding all of that charge at a certain location in a position measurement of the particle. That is what 'charge density' means in the quantum theory of a charged particle. Similarly, the electric *current density* takes the form:

$$\mathbf{j} = -ih(\Psi\boldsymbol{\nabla}\Psi^* - \Psi^*\boldsymbol{\nabla}\Psi) = (\hbar\boldsymbol{\nabla}\alpha)f^2\,, \qquad \text{(I.1.46)}$$

proportional to the same distribution, and in some indirect sense 'proportional' to the momentum which brings in the factor of $\hbar$ and the phase α. We will return to this wavefunction towards the end of this section where we discuss the 'quantization' of charge which can be linked to this particular quantum representation of a particle.

The wave equation for the potentials. Having defined the field strength in terms of the potentials in the equation (II.6.8) , one finds that (in a suitable gauge) the Maxwell equations (I.1.36) reduce to the relativistic *wave equation* for the potentials:

$$\Box A_\mu = \frac{1}{c} j_\mu\,. \qquad \text{(I.1.47)}$$

Also this form of the equations manifestly displays the relativistic invariance of the system: the potentials, and the charge density and current, are neatly organized in four-component relativistic vectors. The wave depicted in Figure I.1.23 corresponds to one of the solutions of the equation (I.1.47) in empty space (without charges or currents).

The solutions of the wave equation are not surprisingly the transversal electromagnetic waves. The wave solution for the gauge potential will look like $A_\mu \simeq \varepsilon_u \exp(i\mathbf{k}\cdot\mathbf{x} - \omega t)$ with the polarization four-vector ε_μ, the so-called *wave-vector* $\mathbf{k}$ and *angular frequency* ω. Substitution in the wave equation shows that we have to impose the condition that $|\mathbf{k}|^2 - \omega^2/c^2 = 0$. The solution corresponds with a wave that propagates in the direction of the vector $\mathbf{k}$, where it has a wavelength equal $\lambda = 2\pi/|\mathbf{k}|$, and a frequency $\nu = \omega/2\pi$. And as expected, the wave condition $\nu = c/\lambda$ is satisfied.

To see the link with the wave depicted in Figure I.1.23, we have to do some more work. First we have to realize that the derivative vector nabla acting on the gauge potential just brings down a factor $\sim \mathbf{k}$ while the time derivative brings down a factor $\sim \omega$. Then we can look at the definitions (I.1.33) to conclude that $\mathbf{B} \simeq \mathbf{k} \times \mathbf{A}$, while we can choose $\mathbf{k}\cdot\mathbf{A} = 0$ which gives $\mathbf{E} \simeq \omega\mathbf{A}$. With these choices we have ascertained that the three vectors $\mathbf{E}$, $\mathbf{B}$ and $\mathbf{k}$ are mutually orthogonal and that indeed the field momentum

is in the direction of $\mathbf{k}$ as $\mathbf{S} \simeq (\mathbf{E} \times \mathbf{B}) \sim \mathbf{k}$. By finally noting that the waves for $\mathbf{A}$, $\mathbf{E}$ and $\mathbf{B}$ are in phase, we have verified all the features of the figure.

This wave equation for the potentials creates the best starting point for the 'quantization' of the electromagnetic field. As we will see later, the A fields are preferred for two reasons. Firstly, if one wants to quantize the electromagnetic field, it is convenient to think of the A_i fields as generalized 'coordinates,' while the electric fields $E_i \simeq \partial A_i/\partial t$ are like the 'momenta' of the field.

It is actually a quite remarkable fact about the Maxwell equations that as equations they survived both the relativity and the quantum revolution. As we will see it is in the interpretation of going from classical fields to those of quantum that the great revolution took place.

Figure I.1.25: *Gauge transformations of the author as Mr Vector Potential.* The pictures illustrate the idea of smooth local transformations. The information content (the person) is the same but the representations or copies are different.

Gauge invariance: beauty and redundancy 

The introduction of gauge potentials naturally leads to the notion of gauge invariance*. In one sense it signals a residual redundancy in the formulation of the theory. This principle is worth exploring as it plays a crucial role in the formulations of all theories that describe fundamental interactions.*

Once you write the equations in terms of the gauge potentials, another fundamental but somewhat elusive property becomes apparent. We have successfully reduced the electromagnetic field from six to four components, by introducing the potentials A_μ, but what we will argue next is that there is still a redundancy in the definition of the system. Whereas giving the gauge potentials yields a unique answer for the physical $\mathbf{E}$ and $\mathbf{B}$ fields, the converse is not true: a given set of $\mathbf{E}$ and $\mathbf{B}$ fields does not uniquely fix the gauge potentials, and this redundancy is called *gauge invariance*.

Let us change the gauge potential by – yes indeed – a *gauge transformation* involving an *arbitrary* function $\Lambda = \Lambda(x,t)$ as follows:

$$A_\mu \to A'_\mu = A_\mu - \partial_\mu \Lambda\,, \qquad \text{(I.1.48)}$$

where Λ is an arbitrary function. If we calculate the transformed fields $\mathbf{E}'$ and $\mathbf{B}'$, we learn that the field components are invariant: $\mathbf{E}' = \mathbf{E}$ and $\mathbf{B}' = \mathbf{B}$, because for any pair of indices μ and ν we have that $\partial_\mu\partial_\nu\Lambda - \partial_\nu\partial_\mu\Lambda = 0$. In other words the contributions of the gauge function cancell.

Let me note that the gauge transformations form a group: they satisfy the group property that two successive transformations, form again a gauge transformation (where $\Lambda = \Lambda_1 + \Lambda_2$).[7] The observable physics, which resides in the $\mathbf{E}$ and $\mathbf{B}$ fields, is independent of Λ, and therefore the theory is said to be *gauge invariant*. In other words, we have

[7] The curious reader may like to jump ahead and look at the *Math Excursion* on groups on page 635 of Part III.

the freedom to choose any convenient function Λ to describe the physics, which is referred to as the freedom to choose a 'suitable gauge.' This choice is useful for example if one needs to construct explicit solutions, but if one has to quantize the electromagnetic field, then this blessing becomes a burden. You could say that the description of the physics in terms of the gauge potentials is elegant but at the same time redundant. It obscures to a certain extent what exactly the real physical degrees of freedom of the (quantized) electromagnetic field are. The wave equation for each of the four components of the vector potential suggests that there are four independent components to the field, yet looking at the electromagnetic wave of Figure I.1.23 we see that in fact it has only two physical components. This further reduction of degrees of freedom from four to two is due to the gauge invariance of the equations.

Gauge symmetry and charge conservation. The Maxwell equation (I.1.36) and the fact just mentioned that the field strength is gauge invariant means that this system is only consistent if the current itself is also gauge invariant. This property can be used to show that the continuity equation $\partial^\mu j_\mu = 0$ follows from gauge invariance. In other words the conservation of charge is a consequence of the gauge symmetry.

A nice way to show this more directly is by noting that the interaction term between the current and the gauge potentials has to be (i) local, (ii) Lorentz-invariant, and (iii) has to give rise to the correct Maxwell equations, which means that it has to be of the form $\int A_\mu j^\mu \, d^4x$. If we now make the gauge transformation (I.1.48), the coupling term acquires an extra term $\int (\partial_\mu \Lambda) j^\mu \, d^4x$, which has to vanish if the theory is gauge invariant. This term can be recast in a convenient form using the following mathematical identity:

$$\int \partial_\mu(\Lambda j^\mu) \, d^4x = \int (\partial_\mu \Lambda) j^\mu \, d^4x + \int \Lambda(\partial_\mu j^\mu) \, d^4x,$$

which is just writing the derivative of a product of two functions as a sum of derivatives on the individual factors and then integrating over space-time. The first term can be integrated to yield the integrand integrated over the three-dimensional boundary of the space-time volume, but on the boundary of space-time we assume the current j_μ will vanish and therefore so does the integrand. And as the integral of zero is zero, the left-hand side of the equation above is zero. This in turn means that the effect of the gauge transformation on the interaction term equals:

$$\int (\partial_\mu \Lambda) j^\mu \, d^4x = -\int \Lambda(\partial_\mu j^\mu) \, d^4x. \qquad \text{(I.1.49)}$$

Now the elegant argument continues by saying that because the gauge function $\Lambda(\mathbf{x}, t)$ can be chosen arbitrarily, and this means that the integral condition has to be satisfied locally, thus we have to require $\partial_\mu j^\mu = 0$ everywhere.

Stated in words, what we have shown is that imposing local gauge invariance requires the current to which the electromagnetic field couples to be conserved locally. This means that net charge can move around obeying the continuity equation, but it cannot just disappear into nothing. This is a not so surprising but vital result that resonates with our earlier observations that the conservation laws of momentum and angular momentum were a consequence of the space-time symmetries being translational and rotational invariance. In that sense one can say that the gauge transformation is like a rotation in a kind of 'internal space' of allowed gauge transformations. This discussion will be taken up in more detail and generality in Chapters I.2 and II.6 where we will have more to say about the geometry of gauge invariance.

A non-local observable: the loop integral of $\mathbf{A}$. Clearly the gauge potentials, as they are gauge-choice dependent, cannot be real observables, the physics resides in the gauge invariant observables being the electric and magnetic fields. These quantities are *local* in that they can be

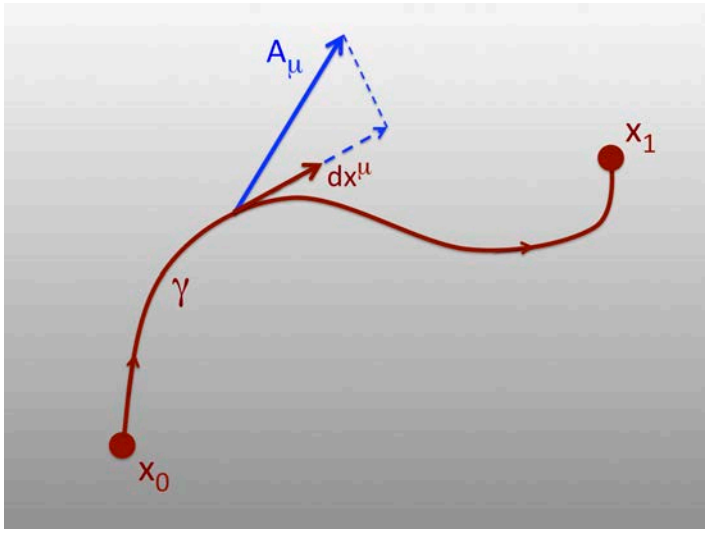


Figure I.1.26: *The line integral of the vector potential* A_μ . The line integral of the four-vector potential A_μ from point x_0 to x_1 along a curve γ . It 'adds' the projections of A_μ along the tangent direction $dx^\mu(\gamma)$ of the curve for all points along γ .

measured locally at a given point x^μ . We may, however, also consider other fundamental gauge invariant quantities, which are intrinsically *non-local* and involve the line integral of the gauge potential A_μ along a closed curve in space-time.

Let us just start by considering a line integral of the vector potential along some curve γ starting at a space-time point x_0 and terminating at point x_1 as depicted in Figure I.1.26 . We write this as follows:

$$I(\gamma; x_0, x_1) \equiv \int_{x_0}^{x_1} A_\mu \, dx^\mu(\gamma) \, . \qquad \text{(I.1.50)}$$

Now let us look what a gauge transformation does to this line integral:

$$\begin{aligned} I(\gamma; x_0, x_1) &\to I'(\gamma; x_0, x_1) = \\ &= I(\gamma; x_0, x_1) - \int_{x_0}^{x_1} \partial_\mu \Lambda(x^\nu) \, dx^\mu = \\ &= I(\gamma; x_0, x_1) - \Lambda(x_1) + \Lambda(x_0) \, . \end{aligned} \qquad \text{(I.1.51)}$$

Clearly the path dependent expression is only affected by the transformation at the start and end point. This implies that if we choose the start and end point to be the same, the resulting 'loop integral' will be gauge invariant as the gauge function Λ drops out. Let us take the example of a closed curve for a fixed time

$$\oint_{\partial D} \mathbf{A} \cdot \mathbf{dx} = \int_D (\boldsymbol{\nabla} \times \mathbf{A}) \cdot \hat{\mathbf{n}} \, d^2S = \int_D \mathbf{B} \cdot \hat{\mathbf{n}} \, d^2S \equiv \Phi \, , \qquad \text{(I.1.52)}$$

where $\hat{\mathbf{n}}$ is the unit vector perpendicular to the surface element $d^2S = dx\, dy$ of the surface D bounded by the curve ∂D . The first equality is an application of the 'Stokes theorem,' which is a mathematical identity explained in the *Math Excursion* on vector calculus at the end of Part III. The second equal sign follows from using the defining relation (I.1.33b) between the vector potential $\mathbf{A}$ and the magnetic field $\mathbf{B}$. Because the contribution of the gauge transformation drops out, this loop integral is gauge invariant and corresponds therefore to a physical and observable quantity, which is not so surprising once you realize that it 'measures' the total magnetic flux Φ through any surface D bounded by the curve, which *is* a gauge invariant quantity.

Gauge versus topological invariance. Yet, there is something quite remarkable about this result. Let us for simplicity consider a two-dimensional plane and have some non-vanishing magnetic flux piercing through the surface area bounded by – say – the unit circle, depicted as the dark region in Figure I.1.27. Outside the unit circle the physical $\mathbf{E}$ and $\mathbf{B}$ fields are zero but that does not imply that the gauge potentials have to be zero there as well. It only requires that the gauge potentials are pure gauge: $A_\mu = \partial_\mu \Lambda$, in other words, that they are a gauge transformation of field $A_\mu = 0$.

Then the result above tells us that you can measure the total magnetic flux Φ through any finite domain, by taking the line integral of the gauge potential around a closed loop which is arbitrarily far removed from that domain. You can

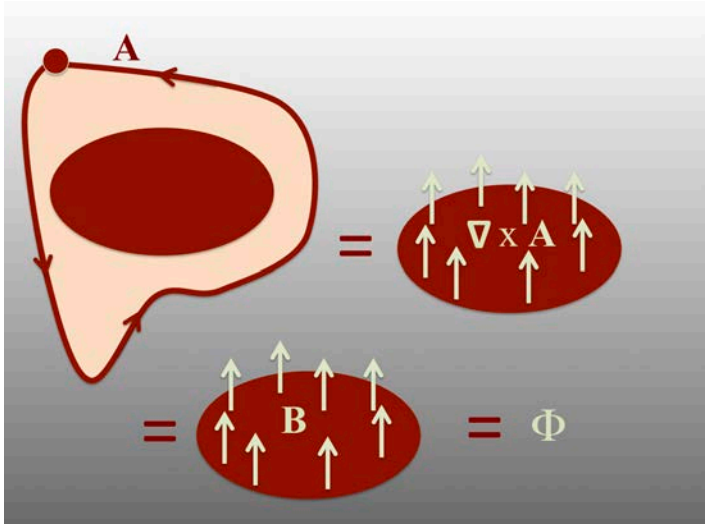


Figure I.1.27: *The loop integral of the vector potential.* A line integral of the vector potential $\mathbf{A}$ along a closed spatial loop is a gauge invariant but non-local quantity. The dark region inside the loop is the region where the magnetic field is non-zero, so, everywhere along the loop there is zero magnetic field, yet the line integral will yield a non-zero magnetic flux Φ .

measure the total flux without ever entering a region where the magnetic field $\mathbf{B}$ is non-zero. Indeed there is a non-local, gauge invariant quantity, corresponding to a measurement outcome that may assume any non-zero value, and that involves probing only a region of space where all physical fields are zero! Quite remarkable indeed!

Imagine we choose the closed loop around a big circle at infinity (the boundary of space) parametrized by $(r = \infty, \varphi)$, then we find for the loop integral simply:

$$I(\gamma = S^1_\infty) = \Lambda(\varphi = 2\pi) - \Lambda(\varphi = 0). \qquad \text{(I.1.53)}$$

Here we run into an apparent contradiction, because on the one hand we argue that the gauge function has to be single-valued meaning that the right-hand side of the above equation should vanish, but on the other hand the left-hand side of the equation is nothing but the loop integral (I.1.52) which equals the total flux Φ !

The resolution of this paradox lies in the appreciation of what we precisely mean by a gauge transformation. We keep the definition simple: a gauge transformation $\Lambda(\mathbf{x}, t)$ is a smooth, single-valued function. Indeed, under such a transformation the value of the loop integral (I.1.52) cannot change. The converse also holds true, if we make a transformation that is *not* single valued, we by definition *do* change the outcome of the loop integral and thereby somehow have changed the magnetic field through the loop.

Let us illustrate this by a simple example: imagine somebody tells me that they have chosen $\Lambda(\mathbf{x}, t) = b\,\varphi$, a constant times the polar angle φ. than the loop integral would give a flux $\Phi = \Lambda(2\pi) - \Lambda(0) = 2\pi\, b$, this does *not* correspond to a proper gauge transformation because it is not single valued. Now it is a matter of semantics what you want to call this transformation; some physicists call it a 'singular' gauge transformation, and others call it a 'topologically non-trivial' gauge transformation. Presumably this is intended to emphasize that it looks like a gauge transformation while strictly speaking it is not, since it is singular at the origin of the plane $(r = 0)$ where φ is not well defined. And indeed such a 'transformation' would 'create' a magnetic flux-line through the point (or the line) where $r = 0$.

In Chapter I.2, in the section on the geometry of gauge invariance on page 96 in, we will see that there is a rigorous topological characterization of the values that the loop integral traversing a vacuum region (or ground state region of some medium) can acquire. The physical situation is determined by a mapping of the closed loop (which is topologically equivalent to a circle S^1_φ) in space into the gauge group G. The outcomes are now determined by the number of topologically distinct ways we can do this and that depends on the global structure of the group-space of G.

For the case of electrodynamics where we have quan-

tized charges the gauge group is the phase group which is also topologically a circle S^1_α. The elements can be represented as $g(\alpha) = e^{i\alpha}$. The constraint that follows is that $\alpha(\varphi) = n\varphi$, meaning that if we go around once in real space then we have to go around an integer n times in the gauge group (so that $g(2\pi) = g(0)$. So the distinct classes are labeled by this integer n with $-\infty < n < +\infty$. So in this theory both the electric charges and the magnetic fluxes would be quantized in suitable units. And because this number is fixed topologically, it is extremely robust. It will not change under *any* smooth deformation of the gauge potentials – not just gauge transformations. The winding number n is therefore a conserved quantity under any smooth deformation, but because it is conserved and quantized for a topological reason, it is called a *topological quantum number*.

We will see later that both gauge invariance and topological invariance play a fundamental role in quantum theory. The loop integral we just discussed is an observable quantity that can be measured as a shift in the interference pattern in a double-slit experiment with electrons, and is known as the *Aharonov-Bohm effect* effect, which is examined in Chapter II.3, after the theorists who proposed this experiment. This effect is a special case of a generalization known as the *Berry phase* which we cover in the same chapter. In an entirely different context the topological invariance of the loop integral can also be linked to the all-important feature of the *quantum statistics* properties of different particle types, like bosons, or fermions as we will discuss in Chapter II.5. ■ ■

Monopoles: Nature's missed opportunity? 

Charge quantization and magnetic monopoles. There is a brilliant, rather early use of the gauge invariance and parallel transport arguments we just presented, by Paul Dirac. In a famous 1931 article he boldly proposed the existence of magnetic monopoles, and proved that the mere

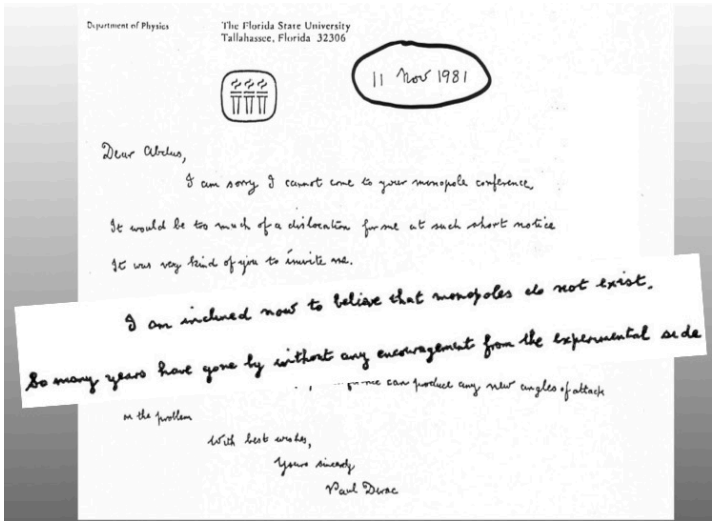

Department of Physics
The Florida State University
Tallahassee, Florida 32306
11 Nov 1981
Dear Abdus,
I am sorry I cannot come to your monopole conference.
It would be too much of a dislocation for me at such short notice
It was very kind of you to invite me.
I am inclined now to believe that monopoles do not exist.
So many years have gone by without any encouragement from the experimental side
with best wishes,
yours sincerely
Paul Dirac

Figure I.1.28: *Dirac in doubt.* This is a fragment from a letter of Dirac to Abdus Salam from 1981, declining an invitation to attend a monopole meeting at the ICTP in Trieste. (Source: Proceedings of Monopoles in QFT, ICTP, Trieste, 1981.)

existence of just a single magnetic monopole in the whole universe would suffice to explain the observed quantization of electric charge! We have already mentioned that a magnetic monopole has never been observed, but that fact by itself does not really exclude the possibility that they somehow exist. May be they once existed and subsequently disappeared through some annihilation process, given that to our knowledge that was what happened to anti-matter, for example. 'To be or not to be,' that is the question, because just being there would suffice!

Dirac's proof goes in fact one way: he shows that if a monopole would exist, then electric charge would have to be quantized in integer multiples of some minimal charge e. In the concluding section of his 1931 article, after noting that the charges we have observed in nature *are* quantized, he modestly states: 'One would be surprised if nature wouldn't have made use of it.'

In practice we can of course do without monopoles be-

cause all magnetic phenomena that have been observed can be explained as being caused by electric currents, moving charges in other words. In all observed magnetic phenomena, there are always a combination of north *and* south poles involved. If you break a bar magnet into two pieces, you get two bar magnets, not a separated north and south pole. And that rule so far holds on all scales, even the smallest accessible. As we mentioned before, this is also the reason that the the sourceless Maxwell equation for the magnetic field reads $\boldsymbol{\nabla} \cdot \mathbf{B} = 0$, where the zero on the right-hand side expresses the merciless verdict: 'No monopoles!' In theory there could have been a 'magnetic' source term there, but there is none. However, the price for it not being there is that the observed quantization of electric charge for the moment remains a mystery. A mystery that has not even been resolved by today's Standard Model of elementary particles and fundamental forces.

The charge quantization puzzle would actually be resolved if the so-called Grand Unified Theories or GUTS turned out to be correct. These theories unify all non-gravitational interactions in one overarching model, as we will discuss in Chapter I.4. This means that different particle types like quarks and electrons belong to a single representation which links their relative charges. Believe it or not, this is precisely the case because those models *necessarily* contain magnetic monopoles in their spectrum as was brilliantly shown by Gerard 't Hooft and Alexander Polyakov independently in 1974. And indeed in these models electric charge *is* quantized. However, these Grand Unified monopoles would be so heavy, of the order of 10^{15} proton masses, that there is no hope making them, even in a fancy lab like CERN. Yet, never say never, may be Dirac will turn out to be right after all. This in spite of the doubt that Dirac himself cast over his prediction towards the end of his life, as expressed in the short note to Abdus Salam depicted in Figure I.1.28.

Dirac's argument. Dirac's argument for charge quantization is sketched in Figure I.1.29. Imagine if we put a monopole with magnetic charge g in the origin, then the magnetic field would point radially outward. The total flux going out through any surface enclosing the monopole is then equal to g. Now imagine the situation sketched in Figure I.1.29, where I draw a sphere around the monopole and I take a charge q and make a closed loop on the surface. Clearly the product of the charge and the gauge invariant loop integral equals the charge times the flux going through the loop. Let me first look at the flux going through the 'northern' surface segment, giving me a flux going upward, say $\Phi_N = \alpha$. However, I could also have taken the flux through the 'southern sector' going down, then that flux would be $\Phi_S = -(g/c - \alpha)$. The phase factors have to be the same (because the flux through any two surfaces bounded by the loop has to be) so we get the following condition on the phases themselves:

$$\frac{q}{\hbar c}\alpha = -\frac{q}{\hbar c}(g - \alpha) + 2\pi n \quad \Rightarrow \quad qg = 2\pi n\hbar c\,. \quad \text{(I.1.54)}$$

Indeed the flux α drops out as it should, because the argument holds for any arbitrary closed loop on the surface. Dirac used the argument exactly the other way around: if there somewhere exists a minimal magnetic charge g, then $qg = 2\pi n\hbar c$. This in turn implies: $q = ne$, so that $e\,g = 2\pi\hbar c$, where e is the minimal electric charge. Therefore he showed that the existence of a magnetic monopole implies the charge quantization that we observe in nature.

Conversely, it is also true that if there existed two particles with incommensurate charges, meaning to say that their ratio would be some non-fractional real number like π or $\sqrt{2}$, then that fact by itself would exclude the existence of magnetic monopoles. So we are left with a stunningly simple and profound explanation of the observed quantization of electric charge, except for the slightly inconvenient fact that we haven't seen any monopole (yet)!

The monopole or Hopf bundle. We have been somewhat cavalier about the precise argument. You could even

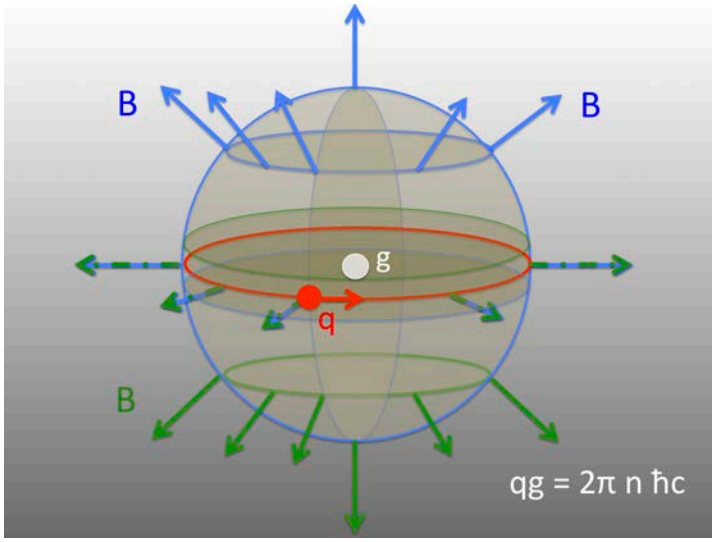


Figure I.1.29: *Electric charge quantization.* This figure illustrates Dirac's 1931 argument for the quantization of electric charge based on the hypothetical existence of a magnetic monopole. To describe the monopole field with potentials requires at least two overlapping patches with potentials $\mathbf{A}_\pm$.

claim that I arrived at the correct answer by incorrect reasoning. You see, the moment I put a magnetic source on the right-hand side of the magnetic Maxwell equation, then it is no longer sourceless. In that case the mathematical identity that $\boldsymbol{\nabla}.(\boldsymbol{\nabla}\times\mathbf{A})\equiv 0$ of equation (A.11a) can no longer hold, which appears to imply that you cannot write the magnetic field in terms of potentials if monopoles are present. Fortunately the situation is not that bad, because the proper use of the potentials turns out to be more subtle. In fact you can still use them, but only locally, as there is a topological obstruction to write a single potential to give the magnetic field everywhere on a surface fully enclosing the monopole. Somewhere on that surface that potential would become singular and the description in terms of a gauge potential would break down. There is a mathematical resolution however, but it is somewhat complicated, and it reveals a fundamental aspect of gauge theories in general. And that is the reason to explore this.

What I am going to to describe you is the mathematical concept of a *fibre bundle*, and we will describe these in more general terms in the section on the 'Physics of geometry' in the next chapter. You could say that we have to enlarge the mathematical framework to that of fibre bundles to allow for situations we couldn't properly cope with before, like having magnetic monopoles.

We start by introducing two coordinate patches S_+ and S_- that cover the sphere, each having the topology of a disc, that have an overlap region with the topology of a cylinder. This is depicted in Figure I.2.30 on page 86 for the sphere S^2, with the blue and green patches S_+ and S_-, and their overlap region containing the equator. Then we define two gauge potentials, say $\mathbf{A}_+$ and $\mathbf{A}_-$ on these two patches that exactly give the magnetic fields present on the patches. So we don't care what $\mathbf{A}_\pm$ do *outside* their patch, they well may develop a singularity *there* but as we don't use them there it doesn't matter. In the overlap region these potentials define strictly identical magnetic fields and therefore have to be related by a gauge trasformation. This is shown in Figure I.1.29, where in the overlap region the two gauge potentials have to be gauge transformations of each other. In terms of equations the statement just made read:

$$\begin{aligned} \text{for } x\in S_\pm \qquad & \boldsymbol{\nabla}\times\mathbf{A}_\pm=\mathbf{B}_\pm\ , && \text{(I.1.55a)}\\ x\in(S_+\cap S_-) \qquad & \mathbf{B}_+=\mathbf{B}_-=\mathbf{B}\,, && \text{(I.1.55b)}\\ & \mathbf{A}_-=\mathbf{A}_+-\boldsymbol{\nabla}\Lambda\,. && \text{(I.1.55c)}\end{aligned}$$

Note that although locally the potentials produce the same magnetic field, what is also clear from the figure is that when we take the loop integral in the overlap region – around the equator for example – then for $e\mathbf{A}_+$, we get the monopole flux through the northern hemisphere $eg/2\hbar c$, but for $\mathbf{A}_-$ we get the flux through the southern hemisphere which has to yield the opposite $-eg/2\hbar c$. This means that the loop-integral over the gauge transformation has to be equal to their difference:

$$\frac{e}{\hbar c}\oint\frac{\partial\Lambda}{\partial\varphi}\mathrm{d}\varphi=\frac{e}{\hbar c}(\Lambda(2\pi)-\Lambda(0))=\frac{eg}{\hbar c}\,. \qquad \text{(I.1.56)}$$

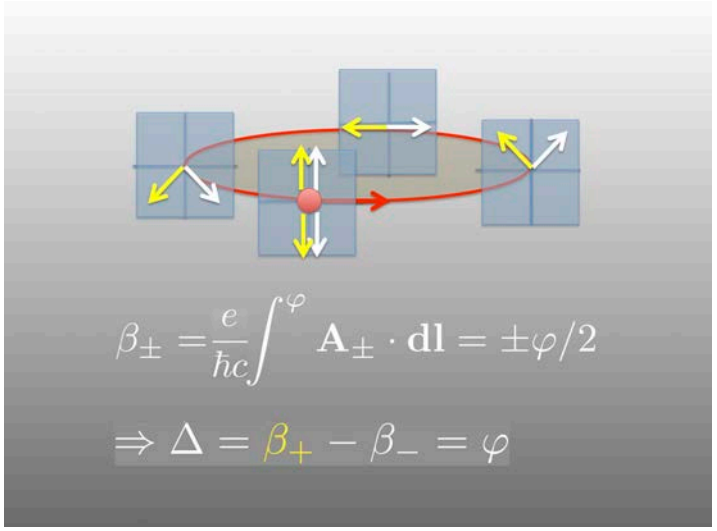


Figure I.1.30: *Parallel transport of charge vector or phase.* around the equator in the monopole field. The phase shifts $\beta^{\pm}$ calculated from $A_{\pm}$ in the northern/southern hemisphere respectively have opposite signs. The requirement for a U(1) bundle is that the transition function $f(\varphi) = e^{i\Delta}$ has to be single valued and has in this minimal case a winding number $n = 1$ because $\Delta(2\pi) = 2\pi$.

If we have a field carrying a charge e it will have an electromagnetic phase factor $e^{i\alpha}$, with a local defined phase $\alpha = \alpha(x)$. This phase will change under a gauge transformation Λ according to $\alpha \to \alpha' = \alpha - e\Lambda/\hbar c$. We may now impose that this charged field is single valued in which case it follows from he eqution (I.1.56) that we to impose $eg/\hbar c = 2\pi n$, the Dirac quantization condition.

In Figure I.1.30 we show an explicit configuration of the phase factors for a charged field. For the elementary $n = 1$ monopole the angles $\beta_{\pm}(\varphi)$ are defined as

$$\beta_{\pm} = \frac{e}{\hbar c}\int^{\varphi} \mathbf{A}_{\pm} \cdot \mathbf{dl} = \pm\frac{\varphi}{2}\,. \qquad \text{(I.1.57)}$$

We learn that the difference between the two line integrals is given by $\Delta = \beta_+ - \beta_- = \varphi$. The loop integrals are gauge invariant and $\Delta(2\pi) = 2\pi$, which means that the transition function $f(\varphi) \equiv e^{i\Delta}$ is single valued.

Topological sectors. The existence of this non-trivial U(1) fibre bundle corresponding to the fundamental monopole with $f(\varphi) = e^{i\Delta(\varphi)} = e^{i\varphi}$ was discovered independently by the German mathematician Heinz Hopf, amusingly in 1931, the same year that Dirac wrote his monopole paper. It took about forty years before the Chinese American physicists Tai Tsun Wu and Chen Ning Yang discovered the mathematical equivalence of these remarkable works of the mind. The bundle space describing the fundamental monopole is basically the three-sphere S^3, and Hopf showed that you can consider S^3 as an S^1 (which equals the group $U(1)$) bundle over a base manifold S^2. We will return to this topological classification of bundles in the next chapter.

So the fibre bundle perspective adds an essential insight into our understanding of electromagnetism as a gauge theory. It is the discovery and classification of topologically non-trivial sectors in the theory. These sectors are defined by mapping of boundaries (or overlap regions) of real space (which themselves are always spaces without boundary) into the gauge group or more generally some 'internal space.' These maps can be non-trivial, and if they are, they label certain topological sectors which define some discrete 'topological charge.' These charges are therefore quantized and conserved for a topological reason which is not directly related to the standard symmetry type argument. Indeed in electrodynamics with monopoles the conservation of electric charge is a consequence of gauge invariance, and the conservation of magnetic charge is topological in nature.

If you look at the monopole as a two-dimensional version of electrodynamics on a closed surface, then the total integral of the magnetic field strength over that closed surface would always have to be an integer in the appropriate units. The total flux is a topological invariant of the gauge field A, because you can make *any* smooth deformation of the gauge field over the surface – not just gauge transformations – and that integer would stay the same. This

total flux which equals the magnetic charge is a topological invariant characterising the gauge field on the surface and is called the *Chern number*. So, indeed, on the two-sphere the discrete values of the magnetic total magnetic flux label different topological sectors of allowed electromagnetic fields. These topological features of gauge theories play an important role in many subfields of physics, for example in understanding the (integer) Quantum Hall effect.

To appreciate the subtlety of the argument let us once more step back and see how it is (quantum) physics that dictates the result. This has to be the case because how else could Planck's constant show up in the charge quantization fomula. That can't be accidental! We see that we map the circle in real space S^1_φ into the gauge group which we was also a circle S^1_Δ. The topological sectors are labeled by the winding number of this map, telling you that $\Delta(\varphi = 2\pi) = 2\pi n$. The compactness of this group tells you therefore two things: (i) that the permitted charges are labeled by integers corresponding to the unitary representations of the group, and (ii) that there are topological sectors corresponding to quantized magnetic charges. If nature had given us particles with arbitrary electric charges like πe or $e\sqrt{2}$ besides e itself, then that would have implied that the gauge group could not have been the compact $U(1)$ but would have been the non-compact group $\mathbb{R}^1$. It's unitary representations are not labeled by integers, so there would be no charge quantization. But at the same time the argument for the existence of non-trivial topological sectors would also collapse. Any mapping of the circle S^1_φ into a line are all contractable to a point, meaning that they are all topologically equivalent, and consequently that there is only one sector in the theory. The world would be without a discrete conserved magnetic charge: no monopoles!

As we will see later the state space of a qubit is also a three-sphere and we will also use the representation of the three-sphere as a bundle space in that context. We will

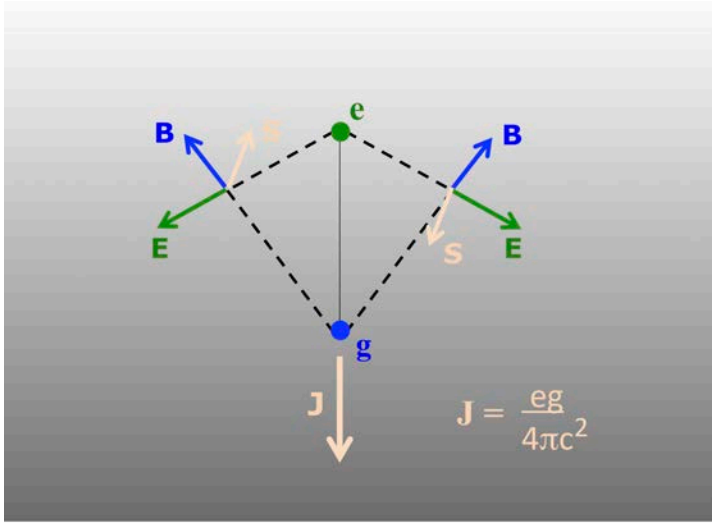


Figure I.1.31: *The charge-pole system.* The charge-pole system is static but has a angular momentum nevertheless. The total angular momentum can be calculated to be equal to $\mathbf{J} = \frac{eg}{4\pi c^2}$, which with the Dirac's quantization condition yields values $\mathbf{J} = (\frac{1}{2}, \frac{3}{2}, \frac{5}{2} \cdots)\hbar$.

return to and expand on these more extended geometrical and topological concepts in Chapter I.2 in the section on *The physics of geometry*. We emphasize these topological features of our mathematical representations of physical systems, because after all topological data refer typically to class labels which are in many cases discrete. In a sense this is a form of quantization that is may be less familiar but certainly no less quantessential. ■ ■

A remarkable case of 'static' angular momentum. 

The system of a spatially separated electric charge – magnetic monopole pair has a curious property first pointed out by J.J. Thomson and presented it as a problem in the Cambridge University Tripod exam in the late 1890s. In Figure I.1.31 we have depicted the situation with the charge and pole located on the z-axis. At two points symmetric with respect to the z-axis we have the electric and magnetic fields $\mathbf{E}$ and $\mathbf{B}$, and the resulting Pointing or field-momentum vec-

tor $\mathbf{S}$. The contribution to the angular momentum around the z-axis is clearly pointing along the charge-pole direction. When we integrate all the contributions, we find that the total angular momentum is non-zero and in fact exactly equal to the quantized product of e and g values in the appropriate units. A static system with a non-zero total angular momentum, a value that is quantized in half integral units and does not depend on the distance between the two sources is remarkable indeed. We will return to these properties in Volumes II en III where we discuss the spin and statistics properties of particles in two dimensions. ■

Statistical Physics: from micro to macro

This section is about macroscopic systems consisting of very large numbers of atoms or molecules and focusses on the link between microscopic and macroscopic behavior, between individual and collective (equilibrium) degrees of freedom. The physics of macroscopic phenomena evidently started as a phenomenologically driven discipline, and it followed the Newtonian approach, by applying analytical methods using differential equations to describe continuous media like gases, liquids and to some extent solids. It lead to a rich variety of equations for thermo- , hydro- and aerodynamics. A crucial turning point came with the acceptance of the molecular hypothesis*, the realization that all forms of matter are made up of tiny molecules. This posed a new challenge, namely to derive and explain all the known macroscopic physics starting from applying basic Newtonian mechanics on the molecular level. As one is not interested in the detailed behavior of the individual atoms, statistics serve as a powerful bridge between the incoherent individual dynamics and the often perfectly coherent dynamics of the collective. This led to a fundamental branch of theoretical physics called* statistical mechanics*, which is considered the third great achievement of classical physics. This approach allowed us to understand numerous so-called* emergent *phenomena - the properties of the collective that are not present on the level of the individual atoms. In this section we focus on thermodynamics: we will first give its macroscopic definition and its three basic laws, and then we will show how a statistical physics approach enables a deeper and unified understanding of the subject. The reason why we are focusing on thermodynamics is that it was within that field that the all-important concept of* entropy *as a measure of disorder and information originated.*

Thermodynamics: the three laws

Thermodynamics is a general theory that started with the noble aim to systematically improve the performance of steam engines and the like, but has now also found notable applications for less down-to-earth systems like black holes. A thermodynamical system – think of a fridge or a steam engine, or just an amount of gas in a container – can work and exchange heat or energy with other systems or its environment. Thermodynamics studies the relations between heat, energy and the ability of the system to do work.

Thermodynamics is a macroscopic theory; nowhere does it refer to the specific microscopic structure of the system. However, when we introduce the subject, it is easiest to envisage a simple gas in equilibrium in a container with particular values for the macroscopic state variables, pressure P, volume V, temperature T as depicted in Figure I.1.32. The fourth state variable, the entropy S, is more hidden as it provides a link between temperature and heat as we will see. This system has an internal energy $\mathrm{U}(\mathrm{T})$ which is the total energy of its internal degrees of freedom.

The essentials of thermodynamics are expressed in three

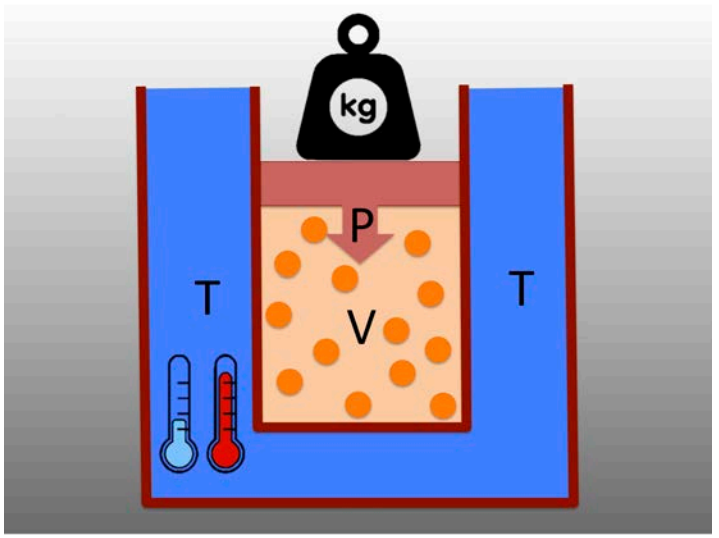


Figure I.1.32: *Gas in thermal equilibrium.* Gas in a container with a movable piston kept at a given temperature T and pressure P, yielding a certain volume V. The three state variables are not independent but satisfy an *equation of state.* For an ideal gas that relation is given in the next figure.

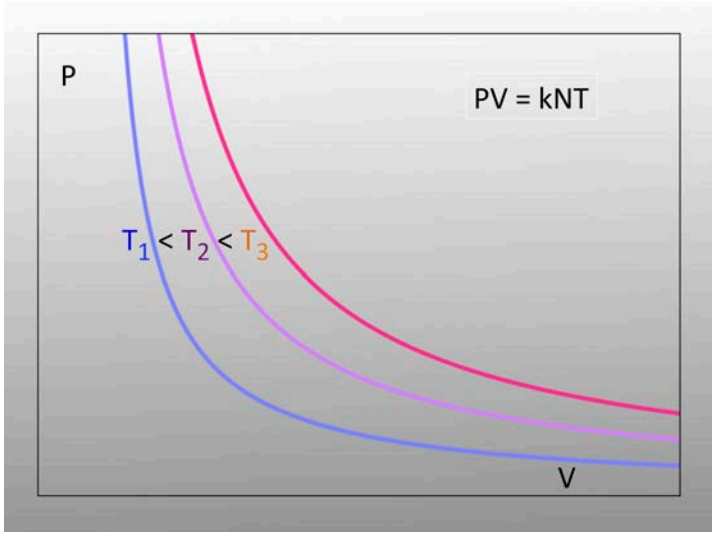


Figure I.1.33: *Ideal gas law.* This graph shows the ideal gas law $PV = kNT$, expressing the dependence between the thermodynamical variables P, V and T, with k the Boltzmann constant and N, Avogadro's number.

famous laws. In fact there is a fourth law, which is usually referred to as the zeroth law of thermodynamics, presumably because it is considered to be self-evident.

The *zeroth law* introduces the notion of thermodynamical equilibrium, and stipulates that it is a transitive property, that is to say that if system A is in equilibrium with B, and A is also in equilibrium with C, then B and C are also in equilibrium. This allows you to define the thermodynamical (absolute) temperature of a system.

The *first law* is basically the statement that energy is conserved. This is expressed in a relation stating that adding some heat dQ to the system will result in an increase of the internal energy dU and the ability for the system to do mechanical work, which for the gas in the container equals the pressure times the change in volume $dW = PdV$:

$$dU = dQ - PdV\,. \tag{I.1.58}$$

The *second law* is the most famous: it features the notion of *entropy*, denoted by S, which is defined by the following relation between heat and temperature:

$$dQ \equiv TdS\,. \tag{I.1.59}$$

This fundamental state variable of any thermodynamical system was introduced by Rudolf Clausius around 1850, as was the second law. The law states that for a closed system (say a fixed quantity of gas in a thermally isolated vessel) entropy can never decrease in time:

$$\frac{dS}{dt} \geq 0\,. \tag{I.1.60}$$

$$\lim_{T\to 0} S = 0\,. \tag{I.1.61}$$

More precisely it goes to a constant which measures the ground-state degeneracy of the system.

Entropy is a sort of measure for disorder: the law boiled down to the familiar phenomenon that (closed) systems

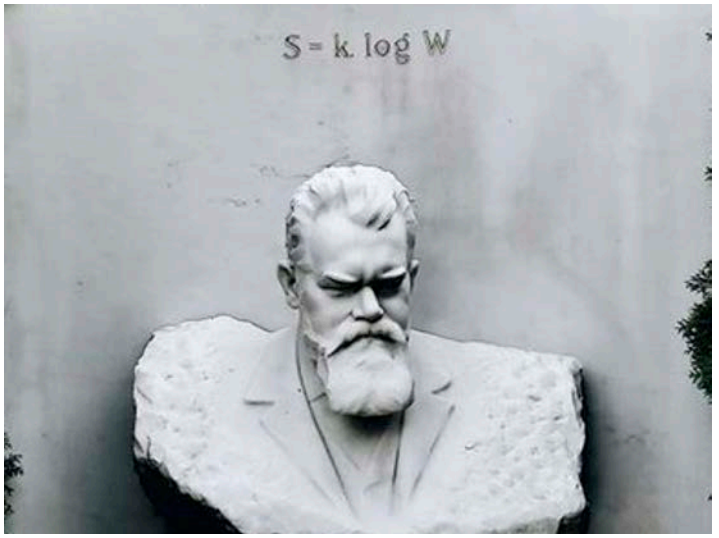


Figure I.1.34: *Ludwig Boltzmann's epitaph.* The expression for the entropy S of a macroscopic state in terms of the number W of microscopic states corresponding to it appears as epitaph on Boltzmann's tomb stone in Vienna's Zentralfriedhof, where he was buried in 1906. (Source: Wikimedia.)

have a natural tendency to become maximally messy or mixed. This applies to teener rooms as well as to tea particles in a pot filled with hot water. As the entropy reaches its maximum value the system reaches an equilibrium state. So, if we put a droplet of ink in a bowl of water, and neither change the amount of water, nor the temperature, nor the volume, then we still see the distribution of the ink molecules through the water changing. In this process the entropy increases until the ink is completely mixed and distributed uniformly and equilibrium is reached. So increasing entropy, you could say, is linked to this process of increasing 'disorder.' To get a deeper understanding of the entropy concept it is necessary to include the microscopic structure of the system whatever that may be. This is our next topic.

Understanding entropy.

The fundamental expression for the entropy S of a given macroscopic state was derived by the great Austrian physicist Ludwig Boltzmann, who stated that it is proportional to the logarithm of the number of microscopic states W corresponding to the macroscopic state under consideration. So

$$\mathrm{S} = \mathrm{k} \log W \,, \tag{I.1.62}$$

where k is not surprisingly called the Boltzmann constant. Now $\log W$ is a pure number and therefore k has units $\mathrm{Joule/Kelvin}$. This famous expression was the precursor of the general notion of the *information capacity* of a system as the logarithm of the number of available states, as it was defined by Claude Shannon in his 1948 foundational paper on information theory. Shortly we will generalize the formula as to establish an explicit connection between statistics and entropy. This relation between entropy and information theory will also be taken up again in the section *The physics of information* in Chapter I.2.

Context dependence of the entropy. To illustrate some features of the entropy concept, we start with some examples of pure *configurational entropy*. Take a system of N boys and N girls that can be located in any of $2\mathrm{N}$ positions. If we furthermore assume that the macroscopic observer is pretty much blind and would have no possibility of distinguishing between boys and girls, nor how many people sit at a given position. So there is no constraint on the configurations and there is only a single macro state. In this case the question is to count the number of possible configurations of $2\mathrm{N}$ people on $2\mathrm{N}$ positions. Now we have to specify the conditions that the micro states have to satisfy. If the people were distinguishable (have names) then the number of possible (micro) states would be $W_1 = (2\mathrm{N})^{2\mathrm{N}}$ as any person can be in any of $2\mathrm{N}$ positions. If we assume they are *indistinguishable*, then we count a micro state where two people are interchanged as the same state, for $2\mathrm{N}$ people we have $2\mathrm{N}$ factorial

different orderings that count as one, and the number of configurations is therefore reduced by this number: $W_2 = (2N)^{2N}/(2N)!$. If we are now on the microscopic level, we could add the distinction between boys and girls, we can exchange the same gender only, and we have to replace the $2N!$ with the much smaller number $N!\,N!$, yielding $W_3 = (2N)^{2N}/(N!)^2$. Next we may add the constraint of *exclusion* meaning that only one person per position is allowed (they behave like fermions), which for the system at hand means that all positions are taken. With name identification the number of configurations is equal to the number of permutations of $2N$ names given by $(2N)!$. With gender identification only we identify the $N!$ permutations of boys and girls separately, yielding $W_4 = (2N)!/(N!\,N!)$.

The effect of resolution and/or constraints. What this little exercise conveys is that the definition of entropy is very much context dependent. Firstly there is the microscopic context of what the degrees of freedom are that one wants to take into account and what the microscopic restrictions are (like distinguishability, exclusion etc.), and secondly there is the macroscopic context determined by what the macroscopic observer is able to distinguish, resolve, or measure (names, gender, spatial compartments etc). So, in general the system has two levels and the entropy is a quantity that basically relates the resolutions (the set of observables and the precision with which these can be probed) and the constraints that determine which states are accessible at each level, and how these observables at the two levels are related. Again, in the examples given above, (i) there was only a single macroscopic state for any given N, and (ii) on the micro level we saw that more resolution leads to more states, while more constraints lead to fewer accessible states. In that sense within a given closed system, indeed, eliminating a constraint leads to 'more disorder' and also a larger number of accessible micro states and thus to an increase of the entropy. In the sequence above we have $W_1 > W_2$ (less resolution), $W_2 < W_3$ (more resolution), and $W_3 > W_4$ (adding a constraint).

The common statement that 'higher entropy means more disorder' is actually quite subtle, and to get a better understanding of this question we add one further structural element to the above example.

Maximal entropy. We consider the previous system with $2N$ positions, but take there to be two compartments, with N positions each separated by a gate, and in each position sits one person. The basic interaction is one where two people exchange position. We start with a special (historically determined) initial state or configuration with all the red-haired girls on the left and the blue-eyed boys on the right. The boys and girls have no names, so exchanging two boys and/or two girls does not change configuration. This means that the initial strict gender separated configuration is a unique one: there is only one such state and it has minimal entropy $S = k\ln 1 = 0$. Next we open the gate in the middle and boys and girls start mixing. I am vaguely suggesting that we are talking about a college dormitory complex in the 1950s, say with $N = 10^3$. The level of frustration among students about the gender separation is like a temperature, and when that becomes high enough, the youngsters start jumping the fences everywhere to go coed. Nice analogy, but now you ask me why this college only admits blue-eyed boys and red-haired girls. I haven't thought about a suitable interpretation for this but no doubt there is one. Physicists, I fear, prefer to think of an ideal gas consisting of equal number of red and blue atoms where $N = 10^{23}$. Let us now increase the resolution of the macro observer and assume that they can somehow measure the number n of boys/girls that are in the 'other' compartment, so the macro-states are labeled by n. Now we ask how many microscopic possibilities there are to realize that particular macro state. The question is to distribute $2N$ youngsters over two partitions. Let us start with one boy/girl jumping the fence: the boy and girl can each come from any of N positions, so for state with $n = 1$ we have $N \times N = N^2$ possible configurations (or micro-states). In the second cross-barrier move, the boy/girl has only $N - 1$ positions to come from or go to, which

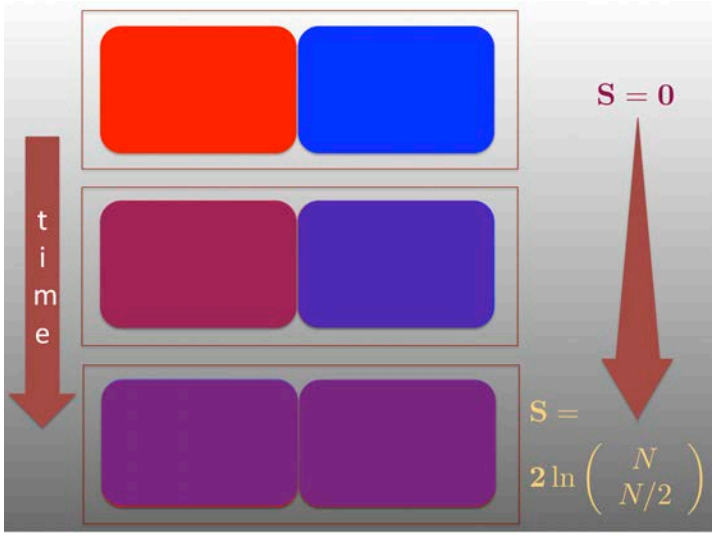


Figure I.1.35: *Gender mixing.* The initial state is the one with all red-haired girls in the left and all blue-eyed boys on the right. When we let them interact through some random girl/boy exchange mechanism, the entropy will increase; equilibrium is reached when the left and right colors have become equal.

means there are $[N(N-1)]^2$ possible $n = 2$ configurations, but now we over-count configurations: we should not have counted the gender neutral exchange of two boys or of two girls as different, so we still have to divide by a factor 4. What we see is that the number of configurations increases extremely rapidly as function of n. The general answer is not too hard to understand from the previous examples:

$$W_n = \left(\begin{array}{c} N \\ n \end{array} \right)^2 \equiv \left[\frac{N!}{n!\,(N-n)!} \right]^2 , \qquad \text{(I.1.63)}$$

where the notation $\left(\begin{array}{c} N \\ n \end{array} \right)$ stands for the binomial coefficient N over n. Note that this function is symmetric under interchange of n and $(N-n)$. Furthermore we observed that the function increases for growing n, then these observations imply that the maximum for W_n is achieved for $n = N - n = N/2$. Thus, if all micro-configurations are equally probable, the macro-state with $n = N/2$ has the largest number of possible micro-states and therefore the largest entropy:

$$S_{max} = k \ln W_{N/2} = 2k \ln \left[\left(\begin{array}{c} N \\ N/2 \end{array} \right) \right] .$$

This means that if we let the random dynamics run for a sufficiently long time from any initial macro-state, and we then probe the system, that we will almost certainly find a configuration with $n = N/2$. So the remarkable insight we gain is that a random process drives the system to a particular macroscopic state, namely the state that is the most probable because it has the most microscopically distinct realizations, which is the state with the highest entropy. And this is the second law of thermodynamics at work. A system has the natural tendency to move from a less to a more probable macro state. That state is the maximally mixed and therefore maximally disordered state that is admissible.

This process is schematically illustrated in Figure I.1.35. To give you a feeling for the numbers involved we have listed the binomial coefficients for various modest values of N and n in Table I.1.1. If N is large, we may approximate the logarithm of a factorial using the famous Stirling formula, which says that, $\ln N! \simeq N \ln N$. With this formula one can show that S in the above equation is well approximated by $S_{max} \simeq 2N \ln 2$, and this in turn implies that in equilibrium the number of micro-states is a number with roughly $0,6\,N$ digits. This implies that the probability for finding the completely gender-separated initial macro-state when the system is in equilibrium is of the order of $p_0 = 10^{-0.6\,N}$. If you take into account that N is of the order of Avogadro's constant $\sim 10^{23}$, p_0 is extremely small indeed.

The arrow of time. There is something rather profound going on in this red-blue dynamics. If you look on the micro-scale all the interchanges are equally probable. In fact any move is its own inverse, and therefore the micro dynamics is invariant under time reversal. If you would re-

n	$\binom{50}{n}$	N	$\binom{N}{N/2}$
1	1	10	252
10	1.0×10^{10}	20	184756
20	2.2×10^{12}	50	0.12×10^{15}
25	1.2×10^{14}	100	0.10×10^{30}
30	2.2×10^{12}	1000	0.27×10^{300}
40	1.0×10^{10}	...	...
50	1	N	$\sim 10^{0.3N}$

Table I.1.1: *The binomial coefficients*. We have listed some values of the binomial coefficients $\binom{N}{n}$ to demonstrate the steep increase as a function of n on the left, and the maximum value of the distribution as a function of N on the right.

Figure I.1.36: *Coarse graining a portrait.* We average the color content over larger and larger (overlapping) squares. In the 'blurring' process the image looses resolution and is therefore hiding ever more information content. The entropy is increasing and the process is irreversible. The entropy is a measure for the amount of micro level information that is 'hidden' for the macroscopic observer.

verse the time direction you wouldn't see the difference in individual moves. On the microscopic level time has no direction! Interestingly, if you look at the macroscopic behavior it clearly has a time direction (namely defined by the increasing entropy) that is not an abstract something or other, this is directly observable. From a macroscopic point of view, when N is large, you see the red compartment slowly turning blueish and the blue compartment slowly turning reddish, but the process stops at a point where both halves have acquired the same purple color. So, somehow the system has created its own *arrow of time*, whatever macro-state you start with it will always move towards the uniform purple color distribution with maximal entropy.

This 'coarse graining' mechanism (see Figure I.1.36) lies at the basis of the time arrow in the real world as well, because, as we have shown in the previous sections, both Newton's and Maxwell's equations are time reversal invariant if the interactions are. What this means is that given a solution to the equations, turning the time around, meaning that we replace t by $-t$, also produces a solution (but may be a different one). On a microscopic level playing the film backward would show another, equally acceptable sequence of events, but on a macroscopic level this is not true. If I drop my bowl of yogurt, fruit and granola on the floor, showing that sequence of events in reverse order, it may be hilarious but it is certainly not of this world. Indeed, this elementary example teaches us that the direction of time is *emergent*, since it has everything to do with the relative number of micro-states belonging to a given macro-state. A randomly propagating system tends to move from a less to a more probable state, and reaches equilibrium in the most probable maximal entropy state.

Two cultures. The second law of thermodynamics paradoxically owes part of its fame to the fact that it is so little known. This was poignantly pointed out by the author (and physicist) C.P. Snow in his provocative essay entitled *Two cultures* published in the *New Statesman* in 1954, in which he bitterly complained about the scientific illiteracy of the cultural elite, and where he used the manifest ignorance about the *second law of thermodynamics* (which in his opinion had a cultural importance comparable to the works of Shakespeare) as a criterion to underpin his criticism. Let me say that Snow's intervention on behalf of thermodynamics did not turn Boltzmann into a Shakespeare. Some years later, however, it did at least provoke a strongly worded reaction from the literary critic F.A. Leavis making the mutual incomprehension even more acute. In a remarkable piece of word craft Leavis stated: 'Snow doesn't know what he means, and doesn't know he doesn't know.' 'The intellectual nullity' he added, 'is what constitutes any difficulty there may be in dealing with Snow's panoptic pseudo-cogencies, his parade of a thesis: a mind to be argued with – that is not there; what we have is something other.' 'But what else to expect from a crappy writer like Snow?' 'As a novelist,' wrote Leavis, 'he doesn't exist; he doesn't begin to exist. He can't be said to know what a novel is.'
The sad point about the situation described by Snow is that it has barely changed over the past half century. So don't ask friends to recite the second law in public, your popularity will most probably instantly plummet. □

This being said we should be cautious, in any given system there will be fluctuations where the entropy actually decreases. The micro-dynamics do not preclude such moves, but on average it is not possible.

It is an awesome idea but certainly correct that in the system we just studied, there is a non-zero albeit inconceivably small probability for the system to pass through the same initial state again!

But that was a state with a lower entropy! The existence of such a recurrence time was proven by Henri Poincaré in 1890. A rough estimate for this recurrence time will be of the order $\tau \simeq 10^{\mathrm{N}} = 10^{10^{23}}\, sec$, which is of the order of 10^{22} times the age of the universe. In whatever units you like to express this truly dazzling number, it is evident that this recursion is not an event to just sit-and-wait for!

This amusingly may remind you of the problems that people who have no understanding of statistics and probability encounter. Events, like the spontaneous gender separation under the given random dynamics in our example, is logically not excluded, but it would take for ever! Assigning outrageously large probabilities to events which are logically not excluded but highly improbable is a specialty of so-called *conspiracy theorists*. Indeed, it would take a conspiracy of extreme proportions to realize such super improbable events, like having all air molecules accumulate in one tiny corner of the room, and you dying because of a lack of oxygen.

Statistical mechanics

The molecular hypothesis. A major step forward was the acceptance of the *molecular hypothesis*, implying that all matter is ultimately build up from microscopic, molecular or atomic constituents. One of the strongest protagonists for this hypothesis was Ludwig Boltzmann. For the number of particles in such macroscopic systems the scale is set by the constant of Avogadro of the order of 6×10^{23} the

number of atoms in a mole of some gas,[8] a number that makes even strong people quiver. *This molecular perspective raised the fundamental challenge for physicists to establish an explicit connection between microscopic physics (mechanics and electromagnetism) and the aforementioned macroscopic laws.* The molecules obey the classical laws, and one - pretty naive - way to think about addressing this challenge would be to face the problem head on and try to solve $\sim 10^{23}$ coupled Newtonian equations for the individual particles simultaneously. Hmm, apart from the computational power needed, this doesn't sound like a very smart idea, does it? Particularly since we are not at all interested in the precise behavior of every individual particle.

Statistical approach. A successful approach is the statistical one, where one links the macroscopic properties like pressure, temperature and entropy to certain average properties of the collective of molecules. Indeed, with such huge numbers statistical methods become extremely powerful and precise as any insurance company can tell you. What properties of the molecular collective could be meaningfully lifted to relevant variables at the macroscopic level? These would typically be the conserved quantities like energy, momentum and particle number. The energy is conserved and for a closed system would be just an additive quantity: the energy of the macroscopic system is just the sum of the individual particle energies and their interactions. The total energy is rigorously conserved: in other words, constant.

Open and closed systems. One is not limited to closed systems, and one might also consider an open system that is coupled to an energy reservoir kept at a fixed temperature (also called a 'heat bath'), which means that one allows for energy (heat) flows between the system and the reservoir as we depicted in Figure I.1.32. If we raise the temperature of the reservoir, heat will flow to the system, raising the internal energy and allowing it to do a certain amount of work. And this gives you an idea of how the first law of thermodynamics can be derived from the microscopic laws. In other words, temperature is the external parameter that sets the average energy of the system, and in that sense imposes an external constraint on the system. For the particle number an analogous reasoning holds. Here one may couple the system to a particle reservoir which is kept at a fixed *chemical potential* μ. This potential corresponds to the energy it costs to add one more particle to the system. These considerations can be made very precise and are part of the field of statistical mechanics, developed by physicists like Boltzmann, Maxwell and Gibbs.

Equipartition of energy. One can show that for a system in equilibrium, on average, the energy is equally partitioned over the individual particles, which means that the notion of temperature is linked to the average energy per particle in the system. In fact the correct way to say this is that the energy is equally distributed over the degrees of freedom, where for a system in equilibrium at temperature T, each degree of freedom gets an energy $\langle \varepsilon_i \rangle = kT/2$. A particle in three dimensions has three independent velocity components and therefore three degrees of freedom. Consequently, for a system of N particles the average energy will be $\langle E \rangle = 3NkT/2$.

Phase space representations of a multi-particle system. Imagine we have a gas that consists of N identical particles in a volume V, then there are two distinct phase space representations of the system possible. One is relevant if one wants to study the average single particle properties or (auto)correlations and refers to the one-particle phase space, while the other concerns the distribution over different multi-particle micro-states that correspond to a single macro-state.

[8] As explained in Chapter I.3, a new definition as of May 20, 2019, of Avogadro's number or constant sets it exactly equal to $N_A = 6.02214076 \times 10^{23}$.

γ - *space.* Let us start with the one-particle phase space $\gamma = (\mathbf{x}, \mathbf{p})$, and represent the state of each particle in the system as a point. This yields a certain density of points, corresponding to a distribution $\mathrm{f}(\gamma, \mathrm{t})$.[9] If the system is in equilibrium, then we expect: (i) the particles to be uniformly distributed in ordinary $\mathbf{x}$'-space, (ii) the distribution to be time independent, and (iii) the momentum dependence to be isotropic. This tells us that in equilibrium $\mathrm{f}(\mathbf{x}, \mathbf{p}, \mathrm{t}) \to \mathrm{f}(|\mathbf{p}|)$, which gives rise to the famous Maxwell – Boltzmann distribution, which is a Gaussian distribution in $\mathbf{p}$ space with the exponent equal to minus the energy: $-\varepsilon/\mathrm{kT} = -\mathrm{p}^2/2\mathrm{mkT}$. Why the exponential energy suppression factor? There are two elementary requirements which make this plausible. If for a simple system like an ideal gas where the particles do not interact and are independent, we look at two particles, then the joint probability to find one of them with p_1 and the other with p_2, we would just be the product of the one-particle probabilities: $\mathrm{f}_2(\mathrm{p}_1, \mathrm{p}_2) = \mathrm{f}(\mathrm{p}_1)\mathrm{f}(\mathrm{p}_2)$. In other words the two-particle configuration should then be weighted by the total energy which is the sum of the two energies. This should hold for any partitioning of non-interacting components which means that the exponential factor is the unique answer, because by multiplying two exponentials the exponents add.

Γ-*space.* We can also define the phase space for the whole system, that total phase space is defined as the Cartesian product of the N individual spaces. This multi-particle phase space $\Gamma_{\mathrm{N}} = \{(\mathbf{x})^{\mathrm{N}},\ (\mathbf{p})^{\mathrm{N}}\}$, of N coordinates and momenta, is $6\mathrm{N}$-dimensional, as each particle has three position and three momentum components. This is a very high-dimensional space, and at any given instant the system as a whole is represented by a *single* point in that space. The particles will bounce around which means that the point representing the system will move around in that space and to study the macroscopic properties of the system we would have to consider long-time averages of those properties. Clearly variables defining macro-states, like for example the total energy, define a constraint on the micro-states, which means that these variables will define certain subspaces or strata in Γ. The micro-states in such a domain can be quite different but cannot be macroscopically distinguished.

Ergodicity. A basic assumption of statistical mechanics, called the *ergodic principle*, is that we can replace the time averages of the system with Γ-space averages using the appropriate distribution representing the equilibrium micro-states. The principle is supposed to hold in the *thermodynamic limit*, where time, volume and the particle number go to infinity (keeping $\mathrm{n} = \mathrm{N}/V$ fixed).

In this setting one may with a single equilibrium state of the macro-system associate a stationary distribution of points corresponding to the probability for the different micro--states representing that macro-state to occur. One introduces a weight function $\rho(\Gamma)$ which may depend on external parameters like temperature or particle number that represent the macroscopic conditions one imposes. Now $\rho(\Gamma)$ defines what is called an *statistical ensemble* of micro-states. If the system is closed (fixed total energy), we speak of the *micro-canonical ensemble*. if we couple it to a thermal bath, we have the *canonical ensemble* with weight function, $\rho(\Gamma) = e^{-\mathrm{H}(\Gamma)/\mathrm{kT}}$, where $\mathrm{H}(\Gamma)$ is the energy function (Hamiltonian) for the multi-particle system. If we also let the number of particles N vary, we get the *grand canonical ensemble*. It was the American physicist Josiah Willard Gibbs who introduced the notion of an 'ensemble' of micro-systems, and the 'ensemble distributions,' to calculate the desired averages in all types of macro-states.

May be to illustrate these rather abstract notions it helps to extend our red-eyed/blue-haired, excuse me red-haired/-blue-eyed youngster model to include variables like 'money' and 'group size.' Clearly the group size is just the number N, we introduced before and we could make it a variable

[9] I refer readers who are not familiar with the basics of probability theory to the *Math Excursion* 'On probability and statistics' on page 626 of Part III.

by coupling to a reservoir of similar pairs who are allowed to join. The amount of money would be the social equivalent of energy, and in a closed system money would be conserved, people could exchange money as long as the total amount of money stays conserved.

If you don't like the analogy, you certainly have a point: whereas in the world of particles there is such a wonderful thing as the equipartition of energy, that is to say that on average every particle has an equal energy, the same does not to seem to hold in the world of money. It's quite the opposite: we witness a process of wealth accumulation. This is a non-equilibrium situation which tends to result in a macabre final state where presumably one person owns all the money. In this case one could speak of the capitalist singularity whereas for the particles one ends up with a socialist uniformity. In this analogy the thermal bath would be represented by the central banks who can raise the fiscal 'temperature' by printing money. I invite the ambitious reader to think about how to include taxation in the model. What these analogies try to convey is that for all these systems there is a notion of a phase space, of external parameters and a statistical ensemble that describes the probability distribution of micro-states depending on an external parameters.

The partition function. The *partition function* of a many-body system is now defined as a phase-space integral, $Z = \int_\Gamma \rho(\Gamma)\, d\Gamma$. You could say that the partition function gives the 'volume' of the domain in $\Gamma-$space, corresponding to the external (macro) parameter choices made in ρ. For example, with ρ describing the canonical ensemble, for a system in contact with a heat bath kept at a temperature T, the partition sum would depend on T as an external parameter.

Emergence. Let us also point out another interesting feature of this statistical approach to systems consisting of many degrees of freedom (particles). In many ways this perspective allows one to introduce 'mean fields' as an approximation to the many body system that underlies it. One passes from a corpuscular perspective to a continuous one. From the macroscopic point of view, a water flow in a river would be described by a mass density field $\rho(x,t)$, a velocity field $\mathbf{v}(x,t)$, and an energy density or temperature field $\varepsilon(x,t)$. These continuous fields are defined by smearing out the local average properties of many particles. You may say that this assumes the existence of a local equilibrium in the system. One may show that these local fields have to obey certain specific dynamical field equations called the laws of hydro-, aero- or plasma-dynamics. These field equations follow from averaging the continuity equations for the locally conserved quantities of the interacting micro system. The resulting laws are '*emergent*' and describe approximately many novel so-called emergent collective properties, in the case of water, you should think of waves and vortices.[10] Water waves are a phenomenon of which the individual water molecules have no idea, the wave property is not present at the constituent level, and it is in that sense that people like to say that the 'whole is more than the sum of its parts.' And it is for that reason that water waves are called an 'emergent' phenomenon. In the simple red-haired-girls/blue-eyed-boys model, we saw the arrow of time emerging, and the emergent (phenomenological) law was telling us that the two colors would uniformly change to the same color purple.

Statistical thermodynamics.

Let us return to thermodynamics. In the statistical approach to a system in *thermal equilibrium*, say, a fixed quantity of gas in a container that we keep at a fixed temperature T, we think of the macro-states labeled by the thermodynamic state variables P, V, S, N and T. In this situation heat can flow from and to the heat reservoir, which

[10] It is striking to see that Maxwell himself believed that his own electrodynamics was an effective description of the collective behavior of an underlying molecular world. As we will see later on, the quantum theory of fields is in a certain way a vindication of that point of view.

means that in thermal equilibrium the energy of the microscopic system is not constant. It will typically fluctuate around the thermal average $U = \langle E \rangle = 3NkT/2$. The relevant energy variable is the (Helmholtz) *free energy* which is defined as:

$$F = U - TS\,, \tag{I.1.64}$$

and should be thought of as a function of T and V, because it follows from the first law that a change in the free energy is given by

$$dF = dU - SdT - TdS = -PdV - SdT\,. \tag{I.1.65}$$

Note that from its definition, minimizing the free energy combines the natural tendencies to minimize the internal energy U and maximize the entropy S.

Let us consider a simple discrete model where each macrostate corresponds to a well-defined set of different configurations on the microscopical level called *micro-states*. This example aims to illustrate how the link between micro- and macro-physics is established. These micro-states are labeled by an index 'i' and each have a certain energy E_i. The probability p_i that a micro-state occurs is again proportional to the Boltzmann weight $w_i = \exp(-E_i/kT)$, which says that the high-energy states are exponentially suppressed.

We may then write that the probability is:

$$p_i = \frac{e^{-E_i/kT}}{Z}\,, \tag{I.1.66}$$

where Z is the *partition sum* defined as

$$Z = \sum_i e^{-E_i/kT}\,. \tag{I.1.67}$$

Note that the sum of all probabilities indeed equals one. The link between the macroscopic and microscopic states is established by giving the expression for the free energy in terms of the partition sum:

$$F = -kT\,\ln Z\,. \tag{I.1.68}$$

From this relation the thermodynamical quantities can be derived. For example with this link it is possible to calculate the famous expression first derived by Gibbs, for the entropy in terms of the probability distribution. Subsequently using equations (I.1.68) and (I.1.66) we obtain

$$F = \sum_i p_i F = -kT \sum_i p_i \ln Z$$

$$= -kT \sum_i p_i(-\frac{E_i}{kT} - \ln p_i) = \sum_i p_i E_i + kT \sum_i p_i \ln p_i\,.$$

Given that by definition $U \equiv \langle E \rangle = \sum_i p_i E_i$, we find from (I.1.64) that the entropy can be expressed as

$$S = -k \sum_i p_i \ln p_i\,. \tag{I.1.69}$$

This is the famous expression for the entropy due to Gibbs which was (re)derived by Shannon, and being the formal definition of information (entropy), forms the basis for information theory. At this point it is important to emphasize the remarkable generality of this result, as it assigns an entropy or information capacity to any given probability distribution or statistical ensemble.

Note that in equation (I.1.69), for a isolated system with fixed energy (not in contact with a heat bath), the energies E_i become equal, and thus $p_i = p = 1/W$. This reproduces the Boltzmann result (I.1.62) for the entropy. Assigning equal probabilities is like saying that you have no a priori information about the states, so you are not imposing any constraint, and thus you get the maximum value for the entropy, the one given by Boltzmann. There is a formal, less physics restricted, method for constructing the maximal entropy distribution as defined in equation (I.1.69) which allows for the systematic inclusion of additional constraints or prior knowledge. This is called the *maximal entropy principle* and is further discussed in the *Math Excursion* 'On probability and statistics' at the end of Part III on page 626. ■

The energy distribution. To further elaborate on the statistical interpretation of thermodynamics, it is illuminating

to look at the energy variable and to derive the energy weight function $s(E)$ from ρ. In the integral over the ensemble of all micro-states, we break the integral up into subsets of equal energy where state i and j belong to the same subset if $E_i = E_j = E$. We call $n(E)$ the volume of a thin shell at energy E.This allows us to write the partition function over all micro-states as

$$\begin{aligned} Z &= \int s(E)\,dE = \int n(E) e^{-E/kT}\,dE \\ &= \int e^{-E/kT + \ln n(E)}\,dE\,. \end{aligned} \qquad \text{(I.1.70)}$$

It is illuminating to go through this calculation for the simple case of an ideal gas, as we will do next.

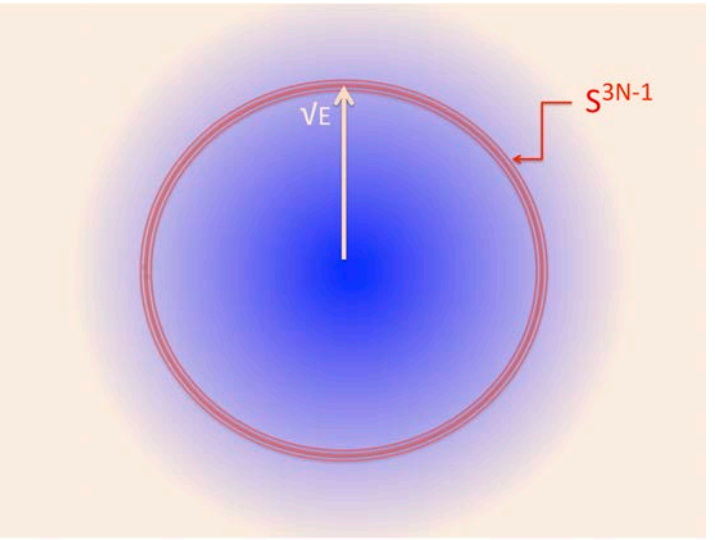


Figure I.1.37: *Phase space distribution.* The rapidly decaying density of points in phase space (in blue). A fixed energy surface (in red) is a very high-dimensional spherical surface. Adding up the points in a narrow shell yields an extremely steeply rising function $n(E)$.

The ideal gas.

Let us consider the ideal gas to show how explicit expressions for the thermodynamical functions in terms of microphysical variables can be obtained by using statistical mechanics. We have N particles in a container with volume V in thermal equilibrium at a temperature T. The total internal energy of a configuration, given by E, equals the sum over one particle kinetic energies: $E = \sum_n (p_n^2)/2m$. To get to the energy distribution we have to integrate (or sum) the general phase space distribution $\rho(\Gamma)$ over all $6N$ variables except the total energy. In an equilibrium state the spatial distribution is uniform and therefore integrating all the coordinates gives a factor V^N. The integral over the $3N$ momenta components has to satisfy the energy constraint that the total kinetic energy equals E. All $3N-$dimensional momentum vectors that satisfy this condition have a length $|p| = \sqrt{2mE}$. So the integral yields the area of a $(3N-1)$-dimensional spherical surface of a $3N-$dimensional ball of radius $R = \sqrt{2mE}$. This means that the density of states takes the form:

$$n(E) = C_N V^N E^{\frac{3}{2}N}\,, \qquad \text{(I.1.71)}$$

where we have dropped a negligible term equal to $1/2$ in the exponent. The constant C_N is the area of the $(3N-1)$-dimensional unit hypersphere.[11]

At this point we can make the connection with thermodynamics, by noting that the entropy of the system is given by:

$$S(E,V) = k \ln n(E) = k \ln C_N + kN \ln\left(V E^{\frac{3}{2}}\right). \qquad \text{(I.1.72)}$$

Solving this equation for the internal energy $U(S,V) = E$ yields:

$$U(S,V) = \frac{1}{V^{\frac{2}{3}}}\, e^{\frac{2}{3kN}(S - k \ln C_N)}\,. \qquad \text{(I.1.73)}$$

If we now use the first law in the form $dU = TdS - PdV$,

[11] The actual expression, which does not enter our considerations, is: $C_N = 3N(\pi)^{\frac{3}{2}N}/(\frac{3}{2}N)!$.

we can determine T and P:

$$\left(\frac{\partial U}{\partial S}\right)_V = T = \frac{2U}{3kN} \tag{I.1.74a}$$

$$-\left(\frac{\partial U}{\partial V}\right)_S = P = \frac{2U}{3V} = \frac{kNT}{V}\,. \tag{I.1.74b}$$

The first equation gives the familiar expression relating the internal energy to the temperature and should be read here as a definition of the temperature in terms of the micro-state energy. From this we may also get the expression for the *specific heat* denoted as c_v, which is the energy needed to raise the temperature by one degree. It is defined as $(\partial U/\partial T)_V$, which in this case yields: $c_v = 3Nk/2$. The second equation gives the *equation of state* for the ideal gas, better known as the ideal gas law $PV = RT$, where the *universal gas constant* R is defined as $R = Nk$. It is an equation of state because it relates the three different thermodynamic state variables P, V, and T. It defines a constrained surface of allowed thermodynamic states, in the space of these three state variables. This basically concludes our first principles derivation of some high-school formulae that apply to the ideal gas.

It is instructive to reflect a bit more on the overall energy weight function s of equation I.1.70. On the one hand, we know that the density of points in the space drops exponentially because of the Boltzmann factor. However, the 'volume' $n(E)$ of the layers grows extremely fast like $E^{3N/2}$, because of the huge value of N. The overall weight, being the product of the two functions, becomes

$$s(E,N) \sim C_N V^N E^{\frac{3}{2}N} e^{-E/kT}\,. \tag{I.1.75}$$

To determine the maximum of $s(E,N)$, we set its derivative equal zero:

$$\left(\frac{\partial s}{\partial E}\right)_N = \left(\frac{3N}{2E_m} - \frac{1}{kT}\right) s(E_m,N) = 0\,. \tag{I.1.76}$$

This yields the value $E_m \simeq \frac{3}{2}NkT = \langle E\rangle$, confirming our expectation that for a very narrow and highly peaked function one expects the maximum and the average to coincide.

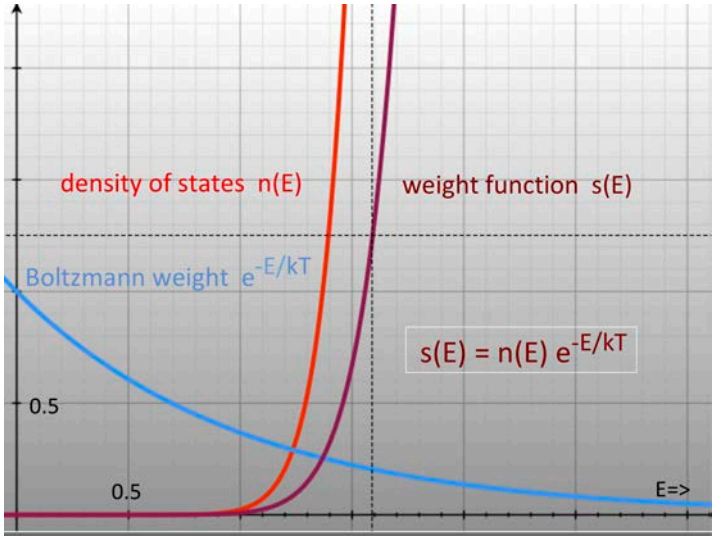


Figure I.1.38: *Energy weight function.* There are three curves, one represents the Boltzmann exponential suppression factor in blue, The density of states $n(E;N)$ in red, and their product, the energy weight function $s(E;N)$ in purple, are plotted near the origin for $N = 8$.

In the Figures I.1.38 and I.1.39 we have illustrated how the resulting weight function $s(E)$ (in purple) emerges as the product of the very steeply rising entropy driven density of states $n(E)$ (in red) and the exponential energy suppression (in blue). We have plotted the case where $N = 8$, which is not quite representative! Indeed it is striking that a narrow peak results: on the left the peak is driven high up by the degeneracy or entropy factor $n(E)$, and on the right it is forced down again by the energy dependent exponential suppression factor. For large N the position of maximum grows proportional to n: $E_m \sim N$, its maximum height increases exponentially: $s(E_m) \sim (const.)^N$, while the width grows only with the square root: $\Delta E \sim \sqrt{N}$. So for large N the relative width decreases like $\Delta E/E_m \sim 1/\sqrt{N}$, and this implies that the weight function becomes proportional to a narrow Gaussian or rather a delta function. And this means that the essential behavior is very well represented by the narrow red band (the hyper-spherical shell) we have drawn in Fig-

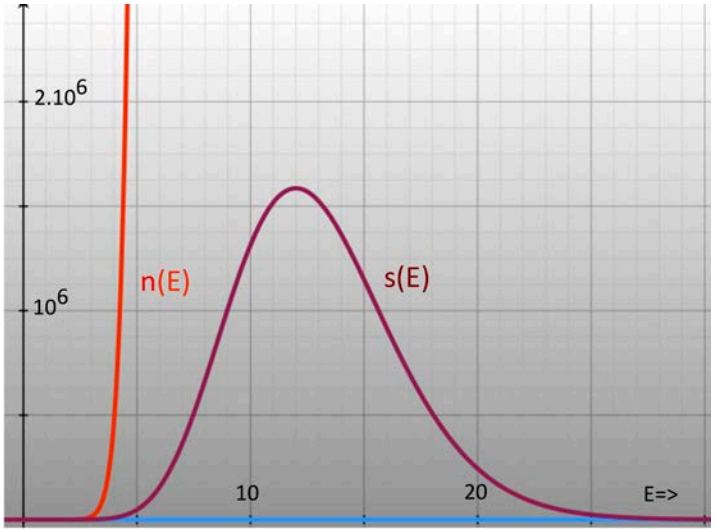


Figure I.1.39: *Ensemble energy weights.* The weight function $s(E;N)$ is the integrand of the partition function. Its maximum increases like $\sim (const.)^N$, the location of the maximum grows $\sim N$, while the width grows only as $\sqrt{N}$. In the limit of very large N, $s(\varepsilon)$ becomes proportional to a delta function. The micro states that matter sit all in an extremely narrow energy band as indicated in Figure I.1.37.

ure I.1.37. Effectively these estimates also show that the energy fluctuations in the canonical ensemble will be very small, which in turn means that effectively the canonical and micro-canonical ensembles are equivalent if we choose $E = E_m$. ■ ■

Classical versus quantum probabilities. We have chosen to highlight this statistical approach to classical many-body physics because we will see that also quantum theory is probabilistic and statistical at heart. And the comparison of the classical statistical physics perspective with the statistical aspects of quantum theory is illuminating. In quantum theory the probabilistic interpretation is forced upon us right from the start at the level of a single particle, and is encoded in the 'wavefunction' description of a quantum particle. The wavefunction is a 'probability amplitude,' and its absolute square represents a distribution. That distribution gives the probability density $\rho(x,t)$ to find the particle at position x at time t, and in that sense it has some mathematical resemblance to the case of statistical mechanics, where the canonical distribution for example gives the probability $p(E_i)$ to find the many body system that has an energy E_i.

There is, however, a fundamental difference between classical and quantum probabilities; in classical physics a probability generally reflects a lack of knowledge about the system, which we in principle could eliminate by making more precise measurements. In quantum physics it reflects a fundamental indeterminism, meaning that even if we have complete knowledge of the quantum state, a property like the spin component along a certain axis for example need not be uniquely fixed. In spite of this difference in interpretation we will see that there are numerous mathematical concepts that can be carried over from statistical physics to quantum theory, and (information) entropy is one of them.

The path integral formulation of quantum theory. From a fundamental perspective a profound yet very direct relation between quantum and classical physics is established through the framework of the (Euclidean) *path integral* formulation of quantum theory proposed by Feynman following an idea of Dirac. The fundamental entity in quantum theory is the probability amplitude A_{if} for the system to go from an initial state labeled i to a final state labeled f. The probability p_{if} for the transition to take place is then given by the square: $P_{if} = |A_{if}|^2$. The amplitude for a quantum particle to go from A at time t_i to B at time t_f can in general be written as a weighted sum over all possible paths $L(t)$ in *classical* configuration space that satisfy the boundary conditions $L(t_i) = A$ and $L(t_f) = B$. As it involves the integration of all possible classical paths or field configurations, the mathematics is quite complicated and in many cases lacks a rigorous mathematical foundation. Yet it is a powerful method that in many ways shows striking mathematical parallels to statistical mechanics if one

makes some substitutions like replacing the energy function with the action functional in the statistical weight, the temperature by the product $\hbar$ and *some coupling*. The notion that the 'free energy' is equal to the log of the partition function translates in the statement that the 'effective action' is the log of the unconstrained path integral over classical configuration space. We return to this topic towards the end of the book in Chapter III.4, after we have gained more familiarity with the quantum world.

Conclusion. Our guided tour along some of the highlights of classical physics has come to an end. To conclude this first chapter, we observe that towards the end of the nineteenth century, many physicists thought that the physical universe was basically fathomed, with only minor details remaining to be settled. The fundamental laws had been laid down by a bunch of geniuses and the program was reduced to merely applying them, skilfully applying them to be sure. That appeared to be a matter of diligent devotion, more something like stamp collecting than facing the challenge of building another Rome in one day...
Indeed, mission almost completed, but as we will see, not quite. Stated differently: Hell was about to break loose!

Further reading.
Some introductory textbooks on classical physics:

- *Classical Dynamics of Particles and Systems*
  S.T. Thornton, J.B. Marion
  Saunders College Publications; 4th edition (1995)
- *Classical Mechanics*
  J.E. Taylor
  Cambridge University Press (2008)
- *Electricity and Magnetism*
  E.M. Purcell
  Cambridge University Press; 3rd edition (2013)
- *Introduction to Electrodynamics*
  D.J. Griffiths
  Cambridge University Press; 4th edition (2017)
- *Fundamentals of Statistical and Thermal Physics*
  F. Reif
  Waveland Press Inc (2008)
- *An Introduction to Thermal Physics*
  D.V. Schroeder
  Oxford University Press (2020)

Complementary reading:

- *The Equations*
  S. Bais
  Harvard Univiversity Press (2005)

Chapter I.2

The age of geometry, information and quantum

> And the continuity of our science has not been affected by all these turbulent happenings, as the older theories have always been included as limiting cases in the new ones.
>
> *Max Born*

In spite of the prevailing scientific optimism towards the end of the nineteenth century, some of the most radical changes in our thinking about the workings of nature were about to surface. The monumental edifice of classical physics started to show cracks which would turn out to be fatal. The crisis in this would-be infallibility centered around some phenomena that were not just hard to explain but were in manifest contradiction with the cherished classical dogmas. The limited domains of validity of classical physics became apparent through the turning points of relativity and quantum theory.
This chapter aims to provide a broad perspective on the new opportunities that opened up for science and technology in the twentieth century, and were derived in some way or another from the turning points that occurred early on. The subsequent sections cover introductions to the physics of relativity, the physics of geometry, and the physics of information. We conclude this chapter with some general remarks on quantum theory.

Canaries in a coal mine

Challenges, contradictions and tuning points. It is interesting to note that already towards the end of the nineteenth century, there were some rather well-known experimental observations that seemed to challenge aspects of the central dogmas of classical physics. We may call these the canaries in the coal mine. Let us start with two results that were puzzling at the time and were only resolved by the radical shift in perspective caused by the theories of relativity, though Einstein himself never emphasized them as sources or motivations for his work. Then we move on to puzzles that pushed us toward quantum theory.

The Michelson–Morley experiment. This experiment succeeded in measuring the effect of the so-called ether (an all-pervading medium through which classical electromagnetic waves supposedly would propagate) on the propagation of light. A non-zero effect was anticipated because the earth would be in motion with respect to the ether and this would cause some dragging of the light in the direction of the relative motion of the ether. Light would therefore propagate at different velocities in different directions. The measurements of Michelson and Morley showed, however, that there was no such effect, leading to the conclusion that the ether was a delusion. It was Einstein who abolished the idea of an ether in his special theory of relativity of 1905.

The (anomalous) perihelion precession of Mercury. It had been observed as early as 1860 that the elliptical orbit of Mercury as a whole rotated very slowly in its orbital plane. This was a problem that even Newton's laws could not account for, even when perturbations like the other planets (even assuming the existence of a novel planet named Vulcanus), as well as the oblateness of the sun were taken into account. But it turned out that the observed anomalous part of the precession agreed to a high precision with the calculation using the *general theory of relativity*, the new theory of gravity formulated by Einstein in 1915. The anomalous perihelion precession thereby furnished one of the earliest experimental confirmations of general relativity.

Let us now turn to four early puzzles that could only be resolved with quantum theory.

The black body radiation law. If we heat a body, it starts to radiate. For a black body kept at a given temperature the classical formula describing the radiation intensity as a function of frequency due to Rayleigh and Jeans failed to describe the data, and in fact predicted an unphysical limit towards the high frequency end of the spectrum referred to as the *ultraviolet catastrophe* (see Figure I.2.1(a)). This all came about because one applied the classical equipartition of energy among the various modes of the electromagnetic field. The resolution of this problem by Max Planck in 1900 was based on the bold assumption that the minimal energy of a mode E is equal to the frequency ν times a fundamental constant denoted by h, according to his famous formula:

$$\mathrm{E} = \mathrm{h}\nu\,. \tag{I.2.1}$$

It is here that the proportionality constant h named after Planck entered physics as a new universal constant of nature. It is extremely small, in ordinary units the reduced Planck's constant – called h-bar – equals

$$\hbar = \mathrm{h}/2\pi = 1.05 \times 10^{-34} \quad \mathrm{Joule\ seconds}. \tag{I.2.2}$$

This tiny constant had a huge impact, since this innocent looking quantization formula marked the very beginning of the tumultuous quantum era.

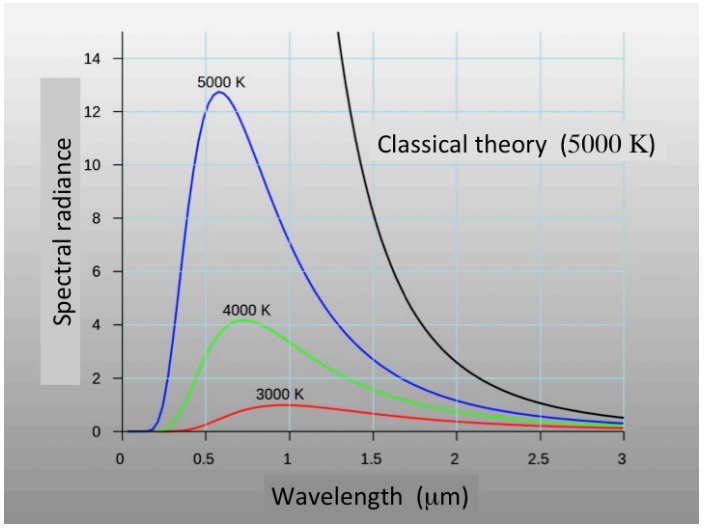
The classical radiation formula can be obtained from the quantum formula by taking the limit where $\hbar$ tends to zero, and in that sense quantum theory clearly marks the limited domain of validity of its classical predecessor.

The structure of the atom. It was known at the time that a gas of atoms of a particular type, like hydrogen or sodium, would absorb or emit light with a specific, discrete spectrum of frequencies. Only narrow lines of particular colors would appear in the spectrum (see Figure I.2.1(b)). Within the classical framework of Newton and Maxwell there was no way to account for this phenomenon, because even accepting the structure of the atom with a positive nucleus and orbiting electrons, there would be no discrete energy levels. Worse still: the electron would radiate and therefore lose more and more of its energy and finally fall into the nucleus. This fundamental instability was basically resolved by Niels Bohr in the quantum mechanical atomic model he proposed, and therefore the stability of all matter we observe is a direct consequence of its quantum nature. Bohr's model for an atom predicted an infinite but discrete set of bound states with a single unique ground state with the lowest energy. And this discreteness accounted for the discrete set of lines in the atomic spectra.

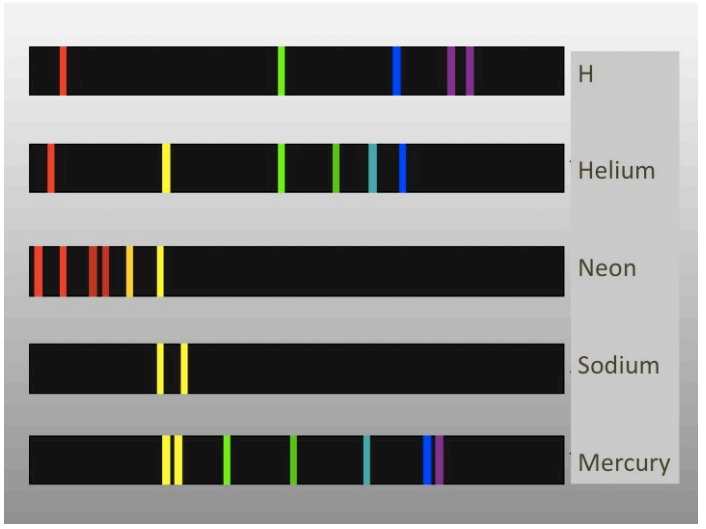
The Compton effect. This effect refers to the fact that when scattering a high frequency X-ray off a charged particle like the electron, the radiation itself behaves much like a particle with an energy E and momentum p given by the Planck formula, i.e.

$$\mathrm{E} = \mathrm{cp} = \mathrm{h}\nu\,. \tag{I.2.3}$$

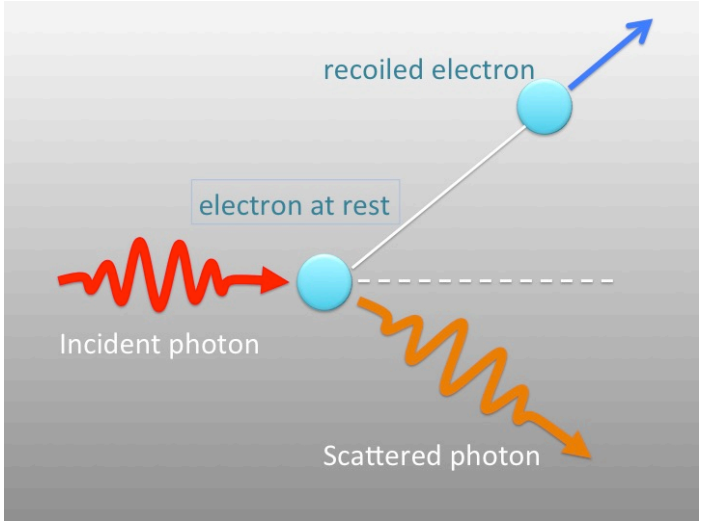
Furthermore, the conservation laws of energy and momentum were respected in such scattering processes (see Figure I.2.1(c)). This clearly suggested the later step made by Einstein who postulated the existence of the *photon* as



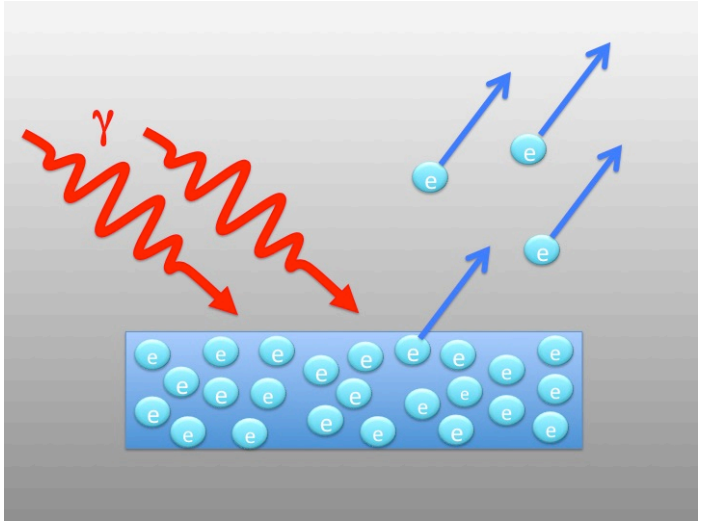
(a) Planck's spectrum of black body radiation solves the ultraviolet (short wavelength) divergence of classical theory.



(b) Discrete lines in atomic spectra, indicating discrete energy levels of the atom.



(c) Compton scattering, showing the particle properties of radiation.



(d) Photo-electric effect with the frequency threshold for the current to flow.

Figure I.2.1: *Meeting the challenge.* Four crucial phenomena that early quantum theory successfully accounted for and where classical physics failed bitterly.

the 'particle of light' with precisely the energy and momentum properties just mentioned.

The photo-electric effect. This amounts to the effect that if we direct a light beam to a metal surface in a constant electric field parallel to the surface, a current might run

because electrons get excited from the surface and flow through the circuit (see Figure I.2.1(d)). The surprise was that the magnitude of the current did not depend on the intensity of the radiation in the way predicted by the classical theory. It turned out that a current would only start running if the frequency of the light in the beam passed a certain critical value. If the frequency was below that threshold, there would be no current irrespective of the intensity of the beam.

This behavior was beautifully explained by Einstein in his 1905 paper, using the particle-like interpretation of the radiation. Only if the energy of a single photon (given by Planck's formula) became larger or equal to the binding energy of an electron in the metal, would the electron be liberated by absorbing the photon. The rest of the energy would be converted into the kinetic energy of that electron.

This concludes our brief summary of some of the deep crises that hit classical physics and that seeded the new paradigms of relativity and quantum theory. These constitute two turning points in our thinking that are unequalled in the history of science in the sense that they extended our understanding of the physical universe far beyond what we as humans could experience and perceive by direct sensing or observation. And to test these radical new ideas many new instruments and experimental techniques had to be developed as powerful extensions of the quite limited innate human ability to probe nature at very small or very large scales. Indeed, these radically new insights started a century of amazing progress, not only in physics and astronomy, but also in chemistry, material science and computer/information science.

The physics of space-time

The theories of special and general relativity, both largely connected with the person of Albert Einstein, showed that there is no objective way to separate time and space, thereby introducing the concept of space-time*. In the special theory of 1905, this implied the unique role of the velocity of light as a universal constant, and the equivalence of mass and energy. The general theory of 1915 furthermore showed that space-time could be curved and had to be thought of as something dynamical. The concept of space-time changed from an external mathematical abstraction to a physical entity, which itself carried energy and momentum. Einstein found the dynamical equations for the universe as a whole, as the inevitable consequence of this line of thinking. This means that we have to think of the universe we live in as a particular solution of the Einstein equations.*

Special relativity

The theory of special relativity is based on two assumptions: (i) the laws of nature should look the same for any set of observers that move with constant relative velocity with respect to each other, and (ii) the velocity of light in vacuum is exactly the same for all such observers. These assumptions, which have been confirmed by a wide variety of precise experiments, have far-reaching implications: for example that the relative velocity between two moving objects can never exceed the speed of light c, but also that moving clocks tick slower. Probably the most well-known consequence is the equivalence of mass and energy, so concisely expressed by the magnificent equation $\mathrm{E} = \mathrm{mc}^2$. This equation opened the possibility of predicting processes, where mass could turn into other forms of energy such as radiation, and the other way around, where for example a high-energy photon could create a particle

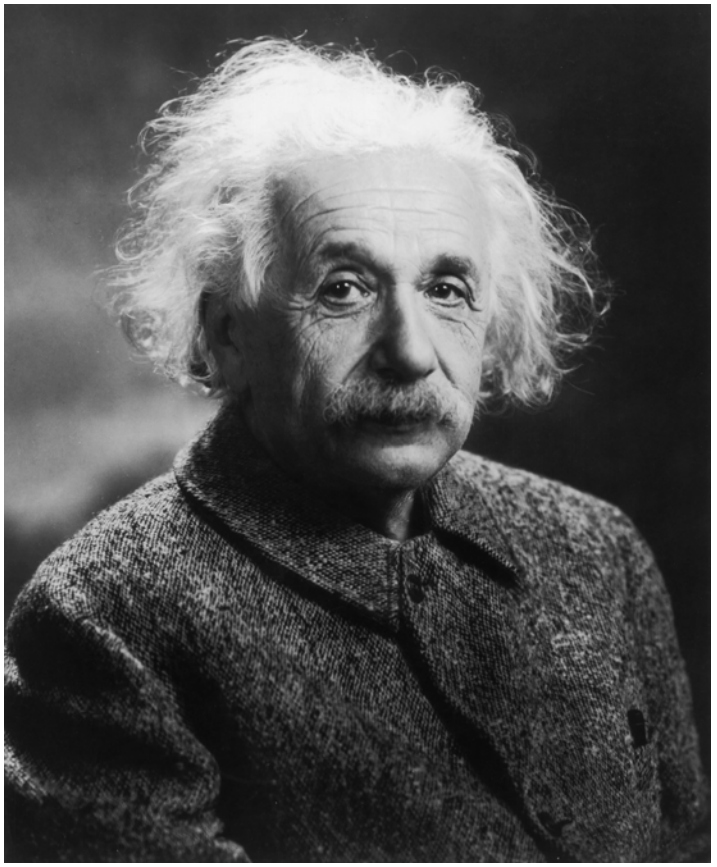


Figure I.2.2: *Einstein.* (Source: Wikimedia.)

anti-particle pair. These processes found ample applications in quantum physics, in particular nuclear and particle physics, as well as in the medical world – think of *positron-electron tomography*, or PET-scanning, as a diagnostic tool.

Space-time four-vectors. From a conceptual point of view Einstein's special theory of relativity introduced the notion of a flat four-dimensional space-time (also called Minkowski space-time) with four coordinates. These are usually denoted by $x^\mu = (ct, \mathbf{x})$ with the index $\mu = 0 \ldots, 3$ where the zero index denotes the time component. A point in space-time labels an instantaneous *event* that takes place at time t at a point $\mathbf{x}$ in space. Correspondingly, Einstein defined a *four-momentum* $p^\mu = (E/c, \mathbf{p})$ for a particle,[1] where the energy became the time component of the four-momentum, with the usual spatial component $\mathbf{p} = m\mathbf{v}$.

If two observers move with constant relative speed, their *four-vectors* that label a specific event, turn out to be observer dependent in a specific way. They would vary, but for the different observers the 'length' of the four-vectors has to be the same. This means that the *space-time interval* s for a given event, defined as $s^2 = c^2t^2 - |\mathbf{x}|^2$ has to be the same for different observers. And similarly, one may define the notion of *rest mass* m_0, for a particle as $m_0^2c^4 = E^2 - |\mathbf{p}|^2c^2$, which is invariant, that is to say that it takes the same value for all relativistically equivalent observers.

The special theory of relativity makes the statement that the physics may look different for different observers, but a complete description can always be given in the frame of any observer. Furthermore, the theory tells you how to calculate what one observer should see if you know the observations from another one. It tells you how to translate any four-vector from one frame of reference to another. And equally important, it also tells you which are the invariant quantities that will be the same in all frames. I emphasize this point about frames here because interestingly enough we will encounter similar challenges if we are to incorporate properties and frames of observers in quantum theory in a consistent way.

Relativistic versus rest mass. Let us dwell a little more on the equivalence of mass and energy. We have so far given two expressions for the energy: one is the canonical $E = mc^2$, and the other $E^2 = m_0^2c^4 + |\mathbf{p}|^2c^2$, involving its rest mass. The latter formula is depicted in Figure I.2.3.

[1] The appearance of the velocity of light c, with units $[m/s]$ in the above definitions, is natural as it ensures that the units of the four components of a relativistic vector are identical.

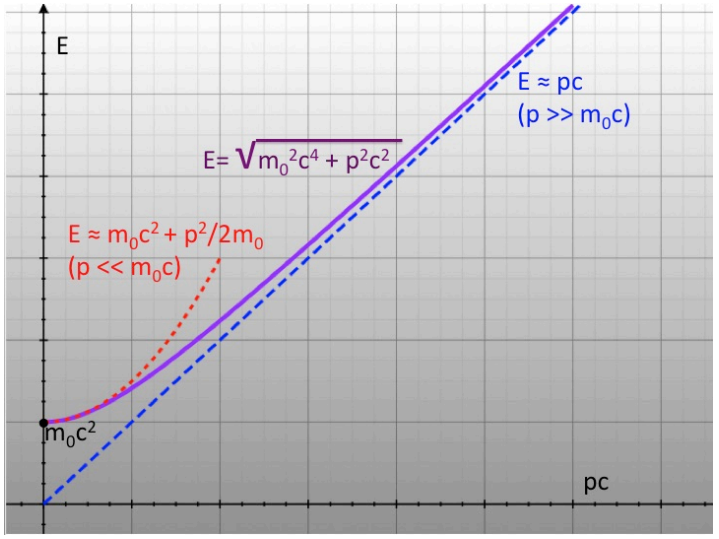


Figure I.2.3: *Relativistic particle energy.* The relation between energy E, rest mass m_0 and momentum p and its limiting behavior for $\mathrm{p} \ll \mathrm{m}_0\mathrm{c}$ and $\mathrm{p} \gg \mathrm{m}_0\mathrm{c}$.

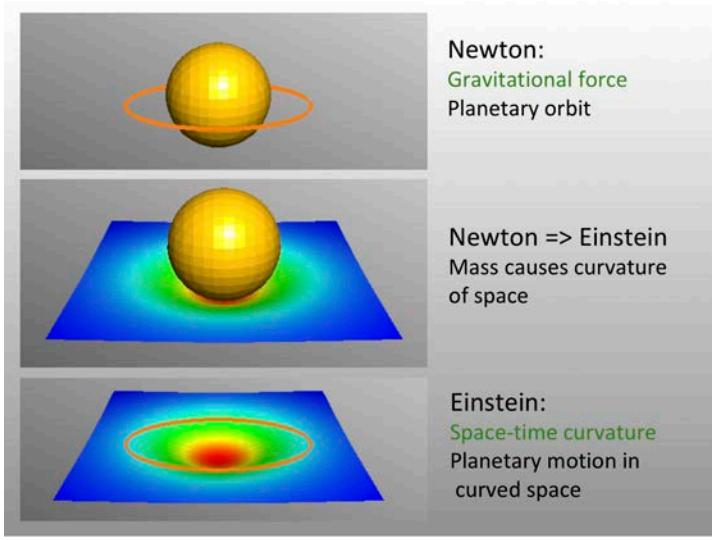


Figure I.2.4: *Mass curves the surrounding space.* Comparing the Newtonian paradigm, where masses cause a gravitational attractive force between sun and planet, and the Einsteinian paradigm where mass curves the space, and the gravitational interaction is induced by way the curved space affects the motion of the planet.

The dashed red curve corresponds to the non-relativistic (Newtonian) limit with $\mathrm{E} = \mathrm{m}_0\mathrm{c}^2 + |\mathbf{p}|^2/2\mathrm{m}_0$, whereas the dashed blue line corresponds to the ultra-relativistic limit where the energy is just proportional with the momentum, $\mathrm{E} = \mathrm{pc}$. Indeed, the latter formula is just the expression for a massless particle like the photon. The picture demonstrates nicely how the properties of a relativistic particle smoothly interpolate between Newtonian particle behavior and a photon. One can also say that the dispersion $\mathrm{E} = \mathrm{E}(|\mathbf{p}|)$ of the particle goes from quadratic to linear.

From the two energy expressions, there follows a relation between the relativistic mass m and the rest mass m_0, reading: $\mathrm{m}^2 = \mathrm{m}_0^2 + \mathrm{m}^2|\mathbf{v}|^2/\mathrm{c}^2$. The conclusion is that in contrast with the rest mass m_0, which is an invariant quantity characterizing the particle, the relativistic mass m is momentum, thus frame and observer dependent. The equation above tells us that $\mathrm{m}^2 = \mathrm{m}_0^2/(1 - \mathrm{v}^2/\mathrm{c}^2)$.[2] If you want to accelerate a particle by applying a force, it is the relativistic mass m that comes in, and therefore particles become effectively extremely massive if their velocity tends to the velocity of light. This in turn implies that to accelerate them further will cost ever more energy. A fact that people who run big accelerators are painfully reminded of every time they receive their utility bills! To be fair I should mention that in an accelerator a large fraction of the energy is lost due to the particles radiating. The relation between masses tells us that the relativistic mass goes to infinity if the velocity approaches the speed of light. No wonder we cannot push particles beyond that universal value!

[2] From hereon we replace $|\mathbf{v}|^2$ simply by v^2 for convenience.

General relativity

The general theory of relativity – often called GR by physicists – is the fundamental theory of gravity proposed by

Albert Einstein in 1915, where the gravitational force is a direct manifestation of the curvature of space-time. In Figure I.2.4 we have indicated the paradigm shift between the Newtonian and Einsteinian perspective on planetary motion. In the Newtonian paradigm the sun and planet have a mass that causes a gravitational force between them, and that attractive force causes the planet to move in an elliptic orbit. In Einstein's picture the masses curve the space around them which is therefore no longer flat. The planet then just feels the curvature of the space it is moving in which causes it to move in an (almost) elliptical orbit. The gravitational interaction is then induced by the curvature of space, like the trajectory of a marble on a rubber sheet deformed by the mass of a heavy bowling ball placed on it. The gravitational interaction manifests itself though the curvature of space-time.

With GR, space-time became a dynamical part of our physical universe. It was lifted from a bunch of silly coordinates to a fully interacting participant. Space-time was promoted from merely a static mathematical arena in which physics unfolded, to a dynamical physical entity, representing physical degrees of freedom carrying energy and momentum itself. You could call it the 'emancipation' from passive mathematical framing to active physical reality.

This development is analogous to electrodynamics, where initially the electromagnetic fields were considered as mathematical constructs that could be used to calculate forces between charges and currents, and only with Maxwell's treatise did it become clear that the fields themselves in a very direct sense represent the physics of electromagnetic radiation. Mentioning this analogy prompts the question of whether a gravitational analog of electromagnetic radiation exists. The answer, as we will see shortly, is affirmative!

General relativity demanded the use of a mathematics that was quite remote from the practicing physicist's repertoire. The language in which gravitational physics was formulated changed from the Newtonian dynamical systems perspective to full fledged Riemannian differential geometry. Relativity marked the beginning of a new golden age of geometry in physics. That is a good reason to include a separate section, following this one, entitled *The physics of geometry*, which provides an introduction to the basic concepts in the mathematics of curved spaces. Concepts that have proven to be as elegant as useful in many domains of modern physics.

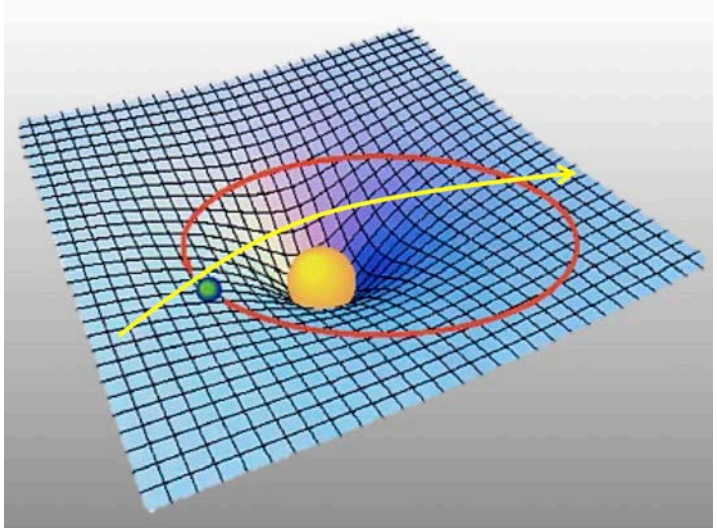


Figure I.2.5: *Bending of light by a mass.* In a curved space-time light moves along shortest distance curves. This means for example that a light-ray emitted by a distant star will be bended if it passes the sun. This effect provided one of the early experimental confirmations of GR.

Seven predictions. The theory of General Relativity made seven almost independent predictions that in the past century have, one after the other, been confirmed experimentally. They are now part of the vast body of experimental evidence supporting the theory. We list them here with a brief explanation:

(i) Bending of light. Generally the geometry of space-time depends on the energy and momentum distribution of radi-

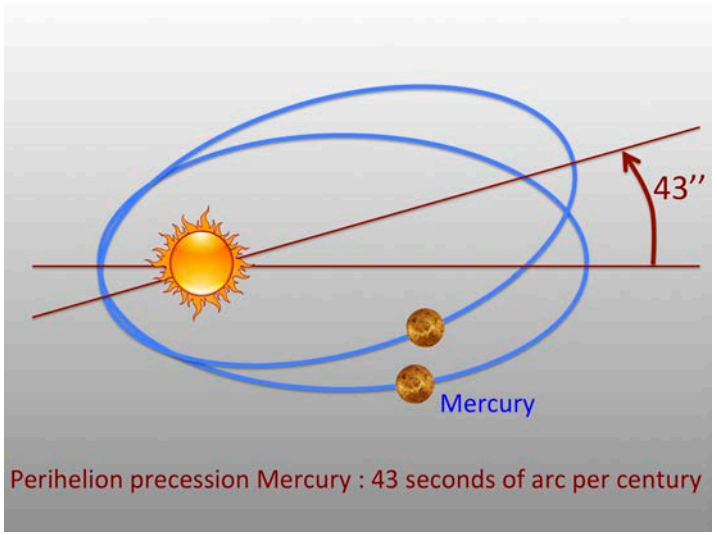


Figure I.2.6: *The perihelion precession of Mercury.*

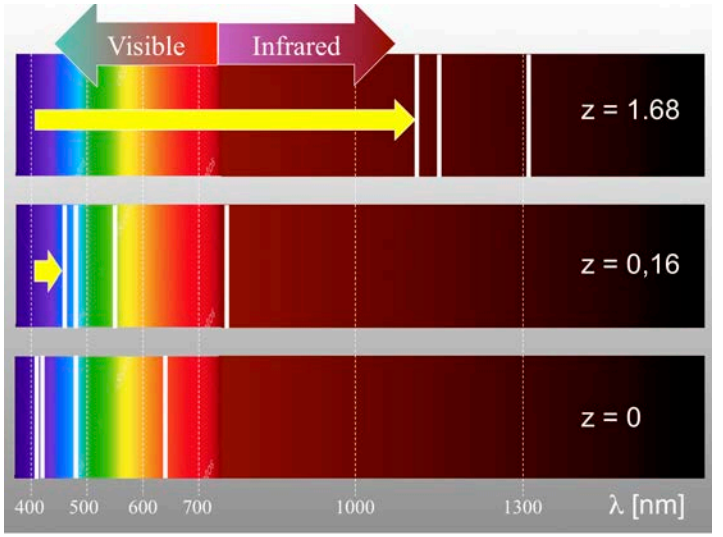


Figure I.2.7: *Gravitational redshift.* If a photon loses energy to the gravitational field moving away from a star its wavelength will increase and gets redshifted. Similarly, due to the expansion of the universe the wavelength of light emitted from far away objects is also shifted towards the red.

ation and matter in it, and in turn that geometry influences the motion of that matter and radiation as through the gravitational force acting on them. We have indicated this effect in Figure I.2.5. This was measured by the British astronomer Sir Arthur Eddington's expedition in 1918 during a solar eclipse, and provided one of the first solid confirmations of Einstein's theory.

(ii) The perihelion precession of planetary orbits. Another notable aspect of General Relativity is that it predicts a deviation from the strictly elliptic orbits for planets. In the Newtonian picture the axes of the ellipse are fixed in space, while in the Einsteinian picture the ellipse rotates slowly in the plane of the orbit as we have schematically illustrated in Figure I.2.6 . One way to understand this is that in General Relativity the effective gravitational force that a static source like the sun exerts on an orbiting planet differs from the Newtonian one. If one expands the potential in powers of $(|\mathbf{L}|/\mathrm{mcr})^2$, one finds that:

$$\mathrm{F} = \frac{\mathrm{GmM}}{\mathrm{r}^2}\left(1 + \frac{3|\mathbf{L}|^2}{\mathrm{m}^2\mathrm{c}^2\mathrm{r}^2} + \dots\right), \qquad \text{(I.2.4)}$$

where m and M are the earth's and solar masses, and $|\mathbf{L}|$ is the angular momentum of the earth. The thing to note is that the Newtonian inverse square law gets a $1/\mathrm{r}^4$ correction. The effect is the largest for the inner planets (small r), for Mercurius the precession amounts to 43 seconds of arc per century. This very slow precession had in fact already been observed before the advent of GR at the end of the 19th century.

(iii) Gravitational redshift. In GR matter and radiation interact with space-time, which means that there will be an exchange of energy between the gravitational and non-gravitational degrees of freedom. So if a photon is emitted from a nearby heavy object like a star and moves radially out to some distant observer, it has to climb out a gravitational potential well and will thereby lose energy. For a single photon this means that the frequency will come down and therefore the wavelength has to increase. The light will therefore be shifted towards the long wavelength or the red end of the spectrum. This effect is called *grav-*

itational redshift denoted by z where the ratio between observed and emitted wavelength is defined by the redshift like $1 + z = \lambda_{\rm obs}/\lambda_{\rm em}$. This gravitational redshift is also predicted to exist for photons coming toward us in an expanding universe, and was the crucial ingredient in the demonstration by Edwin Hubble that our universe is actually expanding. This will be discussed in far more detail in the next subsection.

(iv) Gravitational waves. In a moment we will discuss how these waves were discovered in 2015, exactly one hundred years after their existence was predicted. Gravitational waves are waves in the fabric of space-time that travel with the speed of light. As we have seen the gravitational coupling constant, which is Newton's constant $\mathrm{G_N}$, is extremely small compared to the electromagnetic coupling e. This implies that one needs violent motions of enormous masses to generate gravitational waves that are energetic enough to be detected. For example when black-holes form or collide, there will be huge amounts of energy converted to space-time degrees of freedom. The existence of the waves was one of the early predictions of Einstein's theory, by making a linear approximation to the empty space Einstein equations one does indeed find linear wave equations very much like the equations for electromagnetic waves. It took about a century before this type of radiation was first observed directly on 14 September 2015 by two gravitational wave detectors in the US.

The LIGO project proposed to detect gravitational waves with the use of two giant interferometers. An impressive international effort by the US, the UK, Germany and Austria, that altogether took some 30 years to complete, resulted in the LIGO observatory. Each interferometer takes a laser beam, splits it in two and sends it down two legs at right angles to each other see Figure I.2.9. At the end of each of the legs are mirrors, which bounce the beams back to the center. If there is any difference in the leg length, say caused by the passing of a gravitational wave, the two recombined laser beams create an interference pattern. The LIGO setup was extremely sensitive: it could detect a change in the length of a leg ($\simeq 10^3\mathrm{m}$), on the order of the diameter of a proton ($\simeq 10^{-15}\mathrm{m}$).

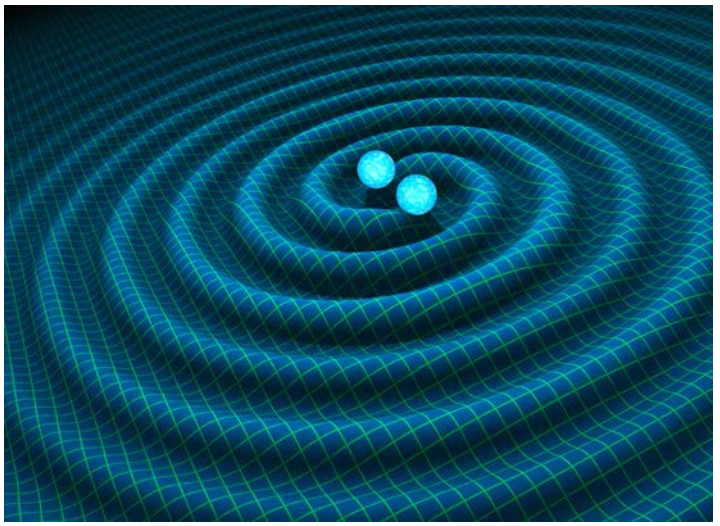


Figure I.2.8: *Two colliding massive objects.* The wavelike space-time profile caused by two extremely massive objects, like black holes, colliding. (Source: LIGO)

The researchers managed to work out the source of the signal, because their model fitted the data so well. Supposedly it was two black holes, 29 and 36 times heavier than the sun merging into a single black hole of 62 solar masses (see Figure I.2.8) meaning that 3 solar masses were emitted in the form of gravitational radiation! As a result of the fundamental importance of the discovery meant that in 2017, the Nobel prize in Physics was awarded to Rainer Weiss, the other half jointly to Barry C. Barish and Kip S. Thorne, 'for decisive contributions to the LIGO detector and the observation of gravitational waves.'

We know that electromagnetic radiation when quantized is directly linked with a massless particle called the *photon*. Likewise, gravitational waves correspond to a massless quantum particle called the *graviton*. As I have said, it couples extremely weakly and therefore will not play any role

Figure I.2.9: *A LIGO gravitation wave detector.* An aerial photograph of one of the two gravitational wave interferometers. A laser beam gets split after which the beams travel forth and back through two orthogonal legs. If a gravitational wave passes through one of them, the two signals show a detectable phase difference after returning. (Source: Advanced LIGO)

in ordinary high-energy accelerator experiments. There is a second fundamental difference between the photon and the graviton: the former is a spin-one particle, and the latter has spin two. This comes about because electromagnetic waves are dipolar, while gravitational waves have a quadrupole moment. In modern views on gravity people tend to think of the gravitational interactions as an emergent phenomenon, which means that Einstein's equations correspond to an effective theory of space-time. It could be that there are more fundamental degrees of freedom (like so-called *superstrings* or *D-branes*) that space-time is really composed of. In that case the quantization of gravity would start from there, and the graviton would rather be a collective excitation, a so-called *quasi-particle*.

The remaining three predictions of GR are:
(v) The existence of black holes,
(vi) The expanding universe,
(vii) A cosmological constant.

These are of fundamental interest in modern physics and therefore we will discuss them separately. The expanding universe and the role of the cosmological constant are the subject of the next subsection on cosmology, while we will discuss some aspects of black holes in the concluding section of next chapter on page 139 .

Big Bang cosmology

The Einstein equations are nothing less than a set of equations for space-time as a whole, which means that our universe should correspond to one of the solutions. These equations have played a glamorous role in 20th century physics and created the astoundingly successful field of observational cosmology. There are many good reasons to present the modern view on the cosmological evolution. It corresponds to the hot Big Bang model described by the Friedmann equation, generalized by Lemaître to include the effect of the cosmological constant. This model describes the dynamical arena in which the world became the way we know it. In the third part of the book we describe in more detail the physical processes that took place at the very early stages of the universe. We will come to appreciate that the combination of understanding basic quantum physics, and cosmology based on GR, leads to an impressive account of the evolutionary process towards an increasing complexity in inanimate matter that preceded the Darwinian biological evolution. Indeed it took the universe billions of years to produce the chemical building blocks of life.

The Friedmann–Lemaître equation. GR in its full generality is quite complicated. However, with a number of simplifying (yet entirely justifiable) assumptions about the structure of our universe, the general equations can be reduced to two strikingly simple equations. The assumptions are referred to as *homogeneity* and *isotropy*, where the meaning of the first is that the universe is the 'same'

at any place at any given instant in time, and the second means that the universe looks the same in any direction at any given instant. And in fact one can show that the second assumption is implied by the first but not the other way around. The first of the resulting equations basically expresses the conservation of energy. The second is the so-called *Friedmann equation*, named after the versatile Russian mathematician and engineer Aleksandr Aleksandrovich Friedmann, who proposed the equation in 1922.[3] The equation reads:

$$\left(\frac{da}{dt}\right)^2 = \frac{8\pi G_N \rho}{3c^2} a^2 - kc^2, \qquad \text{(I.2.5)}$$

where $a = a(t)$ is the *scale factor*, a measure for the relative size of the spatial universe. You may think of a as the relative average distance between two galaxies, meaning that the distance $d(t)$ between the two galaxies at time t would be proportional to $a(t)$: $d(t) = a(t)d_0$. The distances between objects co-moving with respect to the expansion grow proportional to $a(t)$, where in addition we have made the choice that $a(o) = a_0 = 1$. On the right--hand side we have the total *energy density* $\rho = \rho(a)$. Clearly, this is the equation that governs the possible dynamics of homogeneous/isotropic universes. The *'curvature constant'* k, which can be scaled to take the values $1, 0$ or -1, determines whether the space is closed like a sphere, flat, or open like a hyperboloid as illustrated in Figure I.2.10. As we will see the k-value also decides whether the universe will ultimately end in a big crunch ($k = 1$), keeps expanding ($k = -1$), or sits in the critical state ($k = 0$) just in between.

Friedmann sent the equation to Einstein, showing that it had no static solution but did have a solution corresponding to an *expanding universe* originating from an initial singularity. Einstein didn't like the equation, while acknowledging that it was mathematically correct, he thought it was unphysical and 'suspicious' exactly because it predicted

[3] Many physicists also link the names of Lemaître and De Sitter to this law.

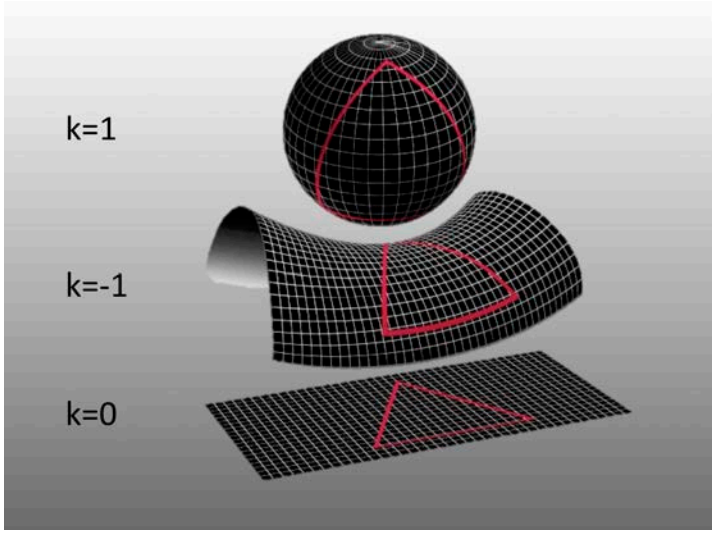


Figure I.2.10: *Curvatures.* The closed, open and flat curvatures corresponding to $k = 1,\ -1,\ \text{and}\ 0$, respectively.

an expanding universe. He then put considerable effort in neutralizing the expansion by adding the so-called *cosmological constant* Λ, without much success. Important work generalizing Friedmann's work including the cosmological constant in 1927 by the Belgian priest and mathematical astronophysicist Georges Lemaître confirmed the expansion.

The real breakthrough came with the mind- and universe-blowing 1929 observations of Edwin Hubble in , which provided the experimental confirmation of the expansion. It was only then that Einstein realized the great importance of Friedmann's work and how he had missed a unique opportunity to make one of the greatest predictions in the history of science. Later in his life he called the introduction of the cosmological constant in his striving for a static universe the 'biggest blunder in my life.' After the expansion was established the new parameter alias cosmological constant silently faded away, until recently when it rather ironically made a glorious and dramatic comeback in a more subtle guise as a term representing the *vacuum* or *dark* energy.

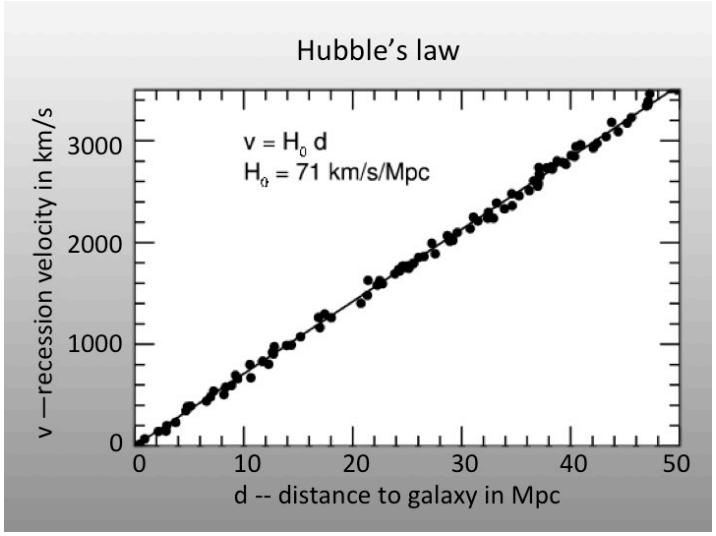


Figure I.2.11: *Hubble law.* Plotting recession velocity of distant galaxies versus their distance gives the linear relation $v = H_0 d$ which is Hubble's law. H_0 is thus the tangent of the angle which the line through the data points makes with the horizontal axis.

The Hubble parameter. An important observable quantity is the *Hubble parameter*, or *expansion rate*, or relative expansion velocity defined as:

$$H(t) \equiv \frac{1}{a}\frac{da}{dt}\,. \tag{I.2.6}$$

So, if we observe the present value H_0 for the Hubble parameter and we determine the total energy density ρ_0 , and put those values back into the Friedmann equation, that would tell us whether k is positive or negative or zero. So in that sense the density determines our destiny. The in-between $k = 0$ case at present ($t = t_0$) defines a *critical density*:

$$\rho_{crit} \equiv \frac{3c^2 H_0^2}{8\pi G_N}\,. \tag{I.2.7}$$

Let me first go back to the definition of the Hubble parameter in equation (I.2.6). If we write it out explicitly for the present time it has a nice interpretation:

$$\left.\frac{da}{dt}\right|_0 = H_0\ a_0 \quad \Rightarrow \quad v = H_0\ d\,. \tag{I.2.8}$$

I read the correspondence as follows: looking from any fixed point in space, I see distant objects at distance $d = a_0$ receding from me with a velocity $v = (da/dt)_0$ then the relation just reads: $v = H_0 d$. This is the celebrated *Hubble law*, and depicted in Figure I.2.11. Clearly the slope in the observed $v - d$ plot gives you the observed value for H_0 . The redshift observations by Hubble in 1929 was one of the great discoveries of 20th century (astro)physics because it implied that our universe was expanding. A fact that – as mentioned – Einstein himself up to that moment did not believe to be possible.

A mechanical analogue. To get a better understanding of the expanding universe we are going to massage the Friedmann equation into a more familiar form, so that we can apply some of our conventional intuitions. Let us first put the constant H_0 back into the Friedmann equation and write it as follows:

$$(\frac{da}{dt})^2 = -H_0^2 \hat{V}(a) - kc^2\,, \tag{I.2.9}$$

where $\hat{V} = a^2\rho(a)/\rho_{crit}$ is some effective 'cosmological' potential. In the modern approach the (relative) energy density has three parts, referring to radiation, matter and the vacuum respectively, thus we write:

$$\hat{V}(\rho, a) = -(\frac{\Omega_r}{a^2} + \frac{\Omega_m}{a} + \Omega_v a^2)\,. \tag{I.2.10}$$

where the omega's are the present values fo the relative energy parameters to be obtained from observation. As I alluded to before, the vacuum term is a remake of Einstein's cosmological constant. It has to be added because other dramatic recent observations have shown that the term is actually there. To understand what all of this means we have plotted the potential for equal values of the Ω's in Figure I.2.12. The qualitative behavior is rather easy to understand: as indicated in the figure, for small a the radiation component dominates, because it comes with the $1/a^2$ factor. For large a it is the vacuum term which dominates as it comes with the a^2 factor. Note that the vac-

uum energy causes an expansive force, it remarkably corresponds to a gravitational repulsion or a negative pressure. The potential is certainly unusual because it has no stable minimum, it runs off to minus infinity, both for a going to zero and for a going to infinity. It is strikingly different from, say, the good old harmonic potential of Figure I.1.13. It is inverted, we have turned it upside-down!

To nevertheless make sense out of it let me remind you of equation (I.1.5) from Chapter I.1, where we derived the expression for the conserved total energy of a particle moving in a potential as:

$$E = \frac{1}{2}mv^2 + V(x)\,, \qquad \text{(I.2.11)}$$

where the total energy E is a sum of the kinetic energy and potential energy $V(x)$. But, lo and behold, that is – up to some substitutions ($m = 2/H_0^2$, $x = a$, $v = da/dt$, and the conserved $E = -kc^2/H_0^2$) – exactly the same as the Friedmann equation (I.2.9).

How remarkable, we have ended up with a one-particle mechanical analogue in 1-dimension for the 4-dimensional universe! That is apparently what cosmic scenarios look like: just kicking a marble and looking at how it is running up and down hill! I don't know who ordered that pizza, but I'll certainly eat it!

The effective cosmological potential $\hat{V}(a)$ looks generically like the dark blue curve in Figure I.2.12. As we have mentioned, this potential has no stable minimum and in fact has two singularities, one at $a = 0$ and the other at $a = \infty$. Apparently there is no fixed scale for the marble-universe to come to rest. Now this is the joy of analogues, they force you to think about what these strange singular features could possibly mean. Cognitive laziness does not suffice, we have to think! Figure I.2.13 shows what the equations are trying to tell us. Well, the singularity at $a = 0$ represents the dramatic event which we called the Big Bang. You could think of it as a marble being shot uphill with considerable kinetic energy so that it can climb the mountain

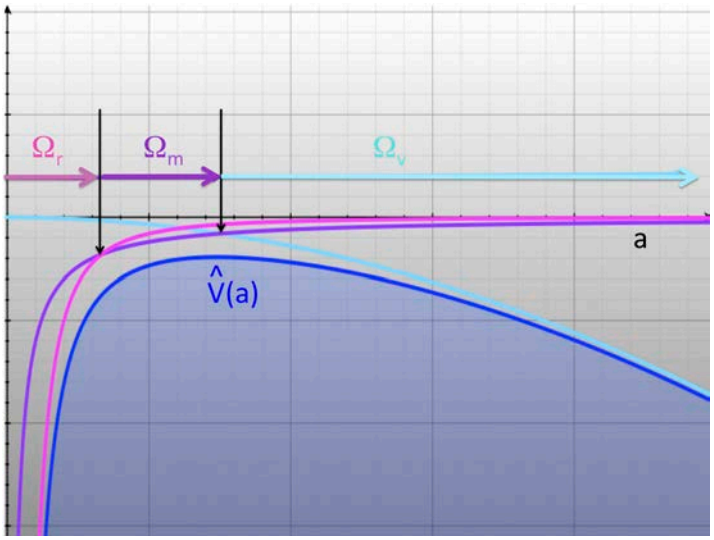


Figure I.2.12: *The effective cosmological potential.* The three terms in the generic cosmological potential $\hat{V}(a)$. The regions in a where the different contributions dominate are indicated (meaning that they are closest to the dark blue curve representing the total potential). For small a radiation dominates, for intermediate scales it is the matter term, while for large values of a the repulsive vacuum term takes over.

from the left. How high? Well, that depends on how hard it gets kicked. If it is kicked a little, it will roll back, and if we slam it hard it will move all the way up, go over the hill and start an infinite descent into another special state. In the latter case the marble-universe keeps accelerating if the vacuum energy density is non-vanishing, causing a race to the bottom on the other side of the potential, a bottom that isn't really there! It describes a state where the universe keeps expanding in an accelerating mode forever, and the matter and radiation will thin out forever with their densities approaching zero.

In Figures I.2.13 and I.2.14 the same three scenarios are depicted: the first shows the potential energy as function of scale factor, and the second the scale factor as function of time. They show three distinct possibilities (with non-vanishing vacuum energy):

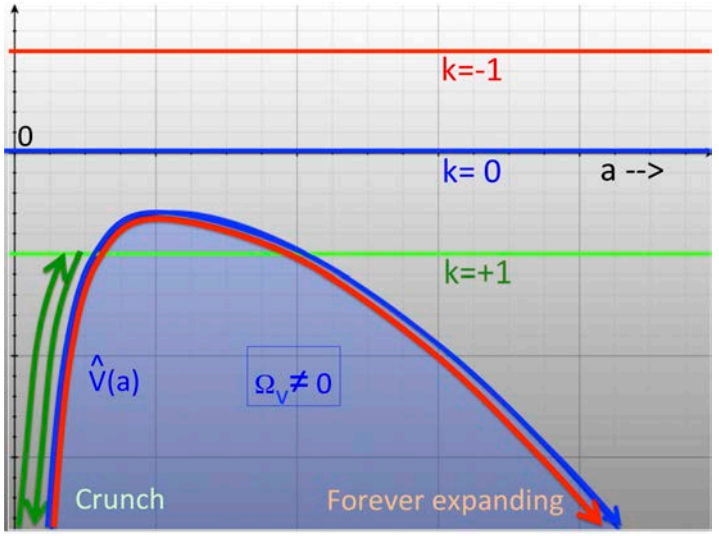


Figure I.2.13: *Evolution scenarios in the potential energy landscape.* The universe with total energy equal to the lines labeled $k = \pm 1$. For negative energy ($\rho > \rho_{crit}$) the evolution follows the green arrows and the universe starts climbing up the potential barrier up to the green line and starts falling back towards a big crunch. If the energy is positive, ($\rho \leq \rho_{crit}$,) corresponding to the red ($k = -1$) and blue ($k = 0$) arrows, the universe easily climbs over the hill and starts accelerating indefinitely.

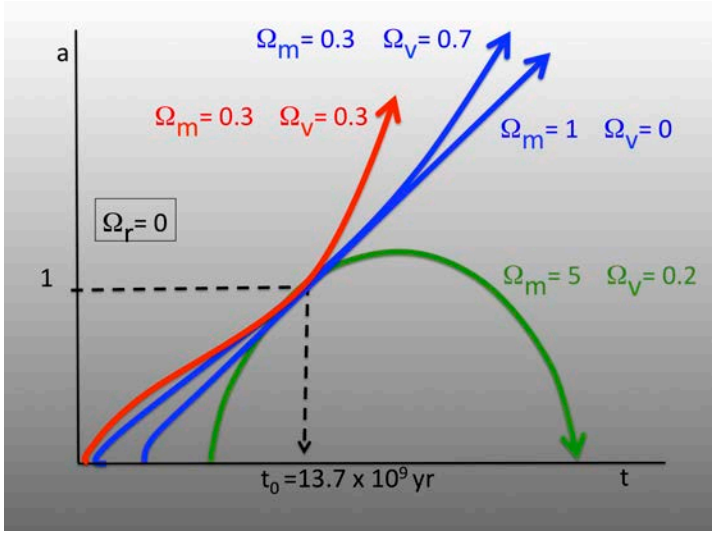


Figure I.2.14: *Cosmological evolution scenarios.* The solutions for the cosmic scale factor a as a function of the time for different choices of the (non-zero) relative mass and vacuum densities. The green scenario is a collapsing universe ending up in a *Big Crunch*. The blue graph on top represents our so-called *Big Chill* universe, it keeps expanding. Compare with the previous figure.

(i) The green scenario with $\rho > \rho_{crit}$ or $k = +1$ is ending in a Big Crunch, because the total energy corresponding to $-k/H_o^2 = -1/2H_0^2$ is not enough to get us over the top. At the point where the marble is turning around, its velocity is zero, which means that all the energy is just potential energy. Consequently the point where the total energy line, corresponding to $k = +1$, intersects with the blue potential energy curve is precisely the turning point of the green arrow that represents the trajectory of the universe.
(ii) In the red scenario with $\rho < \rho_{crit}$, or $k = -1$ the marble moves over the top after which the expansion will go on forever. In this case, there is not enough matter (and radiation) energy to pull the matter back in.
(iii) The $k = 0$ case is of particular interest. If there is a non-vanishing vacuum energy, the top of the potential is at an energy below zero, which means that in the k=0 case the marble still has a non-vanishing velocity at the top and will therefore move over hill entering the domain of eternal expansion.

The Einstein universe. One could imagine cooking up a special case where the top of the potential would exactly touch the $k = +1$ line In this case the marble would end up exactly on the top, where in principle it could stay forever. Forever? But wait, this is like putting a marble on top of a bald head, there is indeed a fixed point, but it is clearly unstable, as any little perturbation will make the marble move one way or the other. In that special case the decision on the fate of our universe would be postponed! The future of the universe would boil down to tossing a coin! The very special solution where the universe just sits forever on top corresponds to the completely static universe that motivated Einstein to introduce the cosmological constant (or vacuum energy term) in the first place. He apparently

didn't check the stability of the solution.

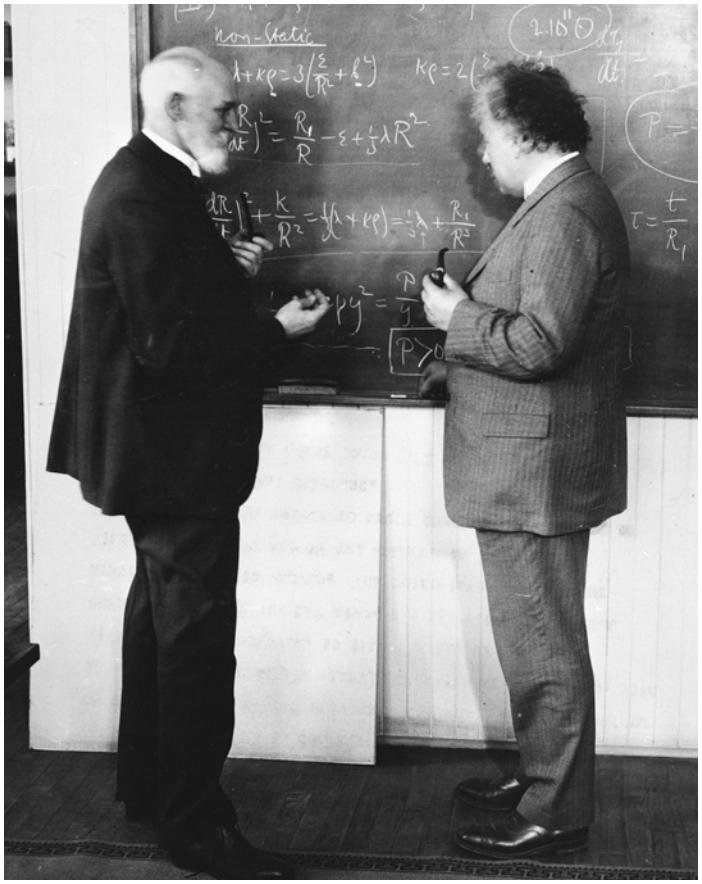


Figure I.2.15: *De Sitter and Einstein in 1932.* De Sitter and Einstein discussing some non-static solutions for the universe. (Source: https://repository.aip.org/islandora/object/nbla:288847.)

Vacuum energy. From the last figures it is also clear that the vacuum energy term is peculiar in that it causes an outward directed force: it acts like a *negative* pressure term. It is apparently a form of energy that gravitationally *repels*! You might be tempted to think of anti-matter, but that can't be it because anti-matter has positive mass, so gravitationally it is attractive like ordinary matter, but this vacuum stuff is peculiar and is really repulsive! In the top-blue and red scenarios we see that for large times the behavior is completely determined by this vacuum contribution, so let us see what happens to the scale parameter in that case. If we go back to the Friedmann equation (I.2.9) and only put in the dominant vacuum contribution ($\Omega_\nu = 1, k = 0$) and bring the a^2 factor to the other side, we get:

$$H(a)^2 = H_0^2\,, \quad \text{or} \quad \frac{1}{a}\frac{da}{dt} = H_0\,. \tag{I.2.12}$$

This equation is simple to solve[4] and yields an exponential expansion:

$$a(t) = a_0 e^{H_0 t}\,. \tag{I.2.13}$$

This exponentially expanding solution is called the De Sitter universe, after Willem de Sitter, the Dutch astronomer who came up with the solution already in 1917. So in the third picture we see the top-blue and red arrow indeed starting to go up exponentially. This solution played an important role in the debates that Einstein and De Sitter (see Figure I.2.15) had about the various non-static universes.

Cosmic event horizon. Expanding universes have the interesting but somewhat puzzling property that if things move away from me at a velocity proportional to their distance, then inevitably at some distance things recede with a faster then the speed of light. This clearly happens as soon as $r > R_H$, where

$$R_H = \frac{c}{H_0}\,. \tag{I.2.14}$$

Can it then be that 'things' move faster than the speed of light? Doesn't that make Einstein turn in his grave? Actually he will not, as his velocity veto concerns relative velocities at a given point in space-time. So indeed, expansion velocities of remote parts of space exceeding c are admissible, and are inevitable in expanding universes. They have a clear physical interpretation, in that they imply the existence of a *cosmological horizon*. In Figure I.2.16 we have sketched the situation. We imagine ourselves to be at the centre with concentric spheres around us. Points on

[4] We will solve it in the *Math Excursion* on functions in Part III.

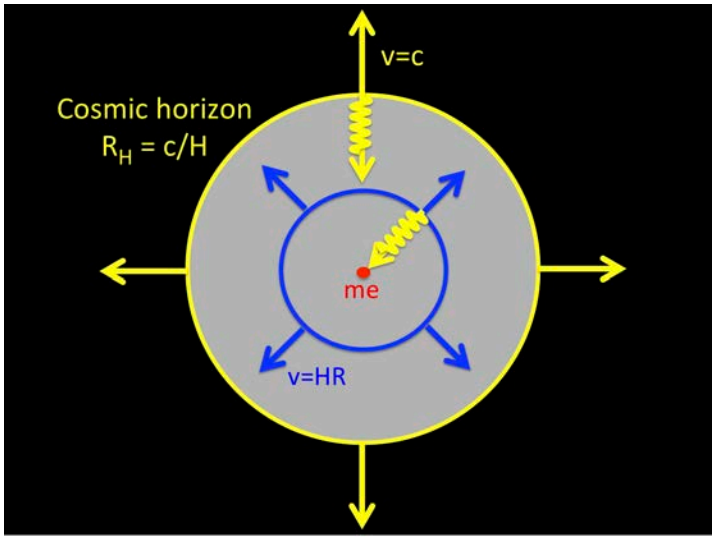


Figure I.2.16: *The cosmic event horizon.* This horizon is defined as the surface around us where the speed due to the expansion equals the speed of light c. Messages sent now from any point in the black region beyond the horizon would never reach us.

the sphere with radius r move away with the Hubble velocity $v = H_0 r$ and the horizon corresponds therefore to the sphere with $r = R_H = c/H_0$. What this means is that if somebody beyond that horizon decides to send us a light signal at this very moment, that signal will never be able to reach us, because it will not be able to cross the horizon. This horizon is at a distance of about 13.7 Giga light years, 'far out' so to speak. Very far away and nothing to worry about. That's what you would think, but after Stephen Hawking's great discovery that horizons have very physical properties: they are a source of thermal black body radiation. Therefore adventurous physicists have been speculating about the conceivable roles this horizon might play in the explanation of contemporary cosmological observations like dark matter and dark energy. We will comment on these ideas later on.

Cosmic inflation

Problems with the standard expansion model.

Particle horizons. We now turn to the phenomenon of a *particle horizon*. This type of horizon should not be confused with the *cosmic event horizon*, as it has a very different origin; the existence of a particle horizon derives from the fact that the universe had a beginning. That means that for any observer at any given instant in time, there is a specific 'particle horizon.' Light emitted from points beyond that horizon never had time enough to reach us. Basically the particle horizon defines the size of the observable universe at any given instant, and the definition naturally implies that the observable universe grows as time goes by. This is schematically illustrated in Figure I.2.17. The particle horizon is just the intersection of our past light cone, with the spatial surface where the time equals zero. This figure also illustrates the notion of a *causally connected domain*, since it has half the radius of the particle horizon. It is the domain in which any point would have had enough time to communicate with any other point in the domain. It is important to note that the younger the universe is the smaller the size of a causal domain. So our observable universe breaks up into ever more causal domains if we go back in time. And this leads to a problem with the standard big bang model and observations that we turn to next.

The (particle) horizon problem. The ever smaller size of particle horizons at earlier epochs of the universe create a notorious paradox known as the 'horizon problem.' This problem concerns a conflict between present-day observations and the original Friedmann-Lemaître expanding model of the universe. We at present observe the cosmic background radiation from all directions in the sky. This radiation was emitted at the moment that electrically neutral atoms formed, when the universe was about 300,000 years old. That radiation did not interact ever since, it de-

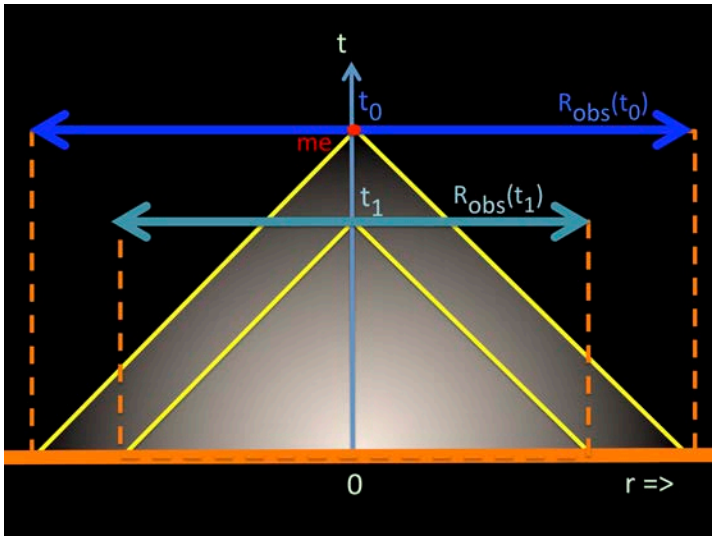


Figure I.2.17: *Particle horizons.* We have sketched the *particle horizon*, which defines the size of the observable universe, at present ($t = t_0$) and an earlier instant ($t = t_1$). It is defined as the present size of the domain that at $t = 0$ was contained in our past light cone, the dark blue arrow. For conceptual clarity the figure features a flat universe with a beginning. It illustrates the fact that our present causal domain breaks up in many independent domains at early times.

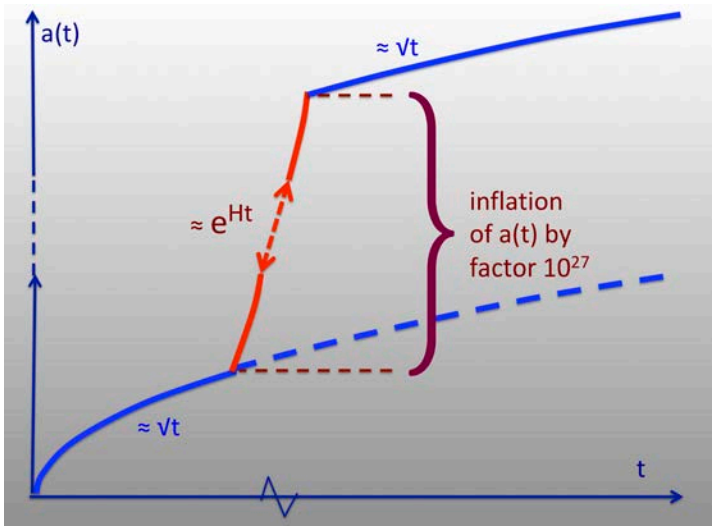


Figure I.2.18: *Causal domains.* The inflationary universe has a brief inflationary epoch of less then 10^{-30} s shortly after the big bang, in which it expanded exponentially with a factor of 10^{27}.

coupled from the matter. And that is the reason why we observe a perfect thermal spectrum now, which is redshifted because of the expansion of the universe after the decoupling took place. It constitutes the strongest direct observational evidence for the expansion of our universe. It appears exactly as predicted. However, there is something puzzling here: the radiation that comes to us from opposite sides of the universe shows exactly the same spectrum apparently originating from the same thermal plasma. How can that be? Because at the time the photons decoupled, the places where that radiation originated were not within one causal domain. To get an idea, let us ook at Figure I.2.17. If we imagine the radiation to be released at $t = t_1$, then it can have equilibrated over distances corresponding to the size of the causal domain with radius $R_{obs}(t_1)$ as indicated in the figure. My causal domain consists of all the points from which information could reach me within the age of the universe t_0. That would mean that at present we would be able to make observations all the way out to $R_{obs}(t_0)$, a much larger domain. Thus we would not expect the perfect black body spectrum we happen to observe. In other words at the time of decoupling the region corresponding to our presently observable universe contained many causal domains. At that age of the universe, there had not been time enough to reach thermal equilibrium over distances that comprise the total observable universe at present. This is an irrefutable fact if we assume that the standard expansion of the universe is correct. And this fact poses a serious problem for the standard Friedmann-Lemaître model. This problem has been resolved by making a major amendment to the course of events in the very early universe, This is a fundamental update: the expanding universe 2.0, also called the *inflationary universe*. But before we get into that we want to first mention another problem with the standard cosmological model.

The flatness problem. The flatness problem is posed by the observation that fitting the model to the data the conclusion is that we live in a universe where the curvature constant k is very close to zero. From a theoretical point of view there is no reason to expect it to be zero, it must have been zero all along. From the fact that our universe after 13.7 billion years has a k value so close to zero, one may show by calculating backward that this would impose a very unnatural initial condition on the universe. One finds that the value of the curvature constant would have to be fine-tuned to zero to some sixty decimal places! That is considered to be an exceptional choice, which begs for an explanation. It turns out that there is a satisfactory solution to this problem and again it involves the vacuum energy and the De Sitter solution.

If you go back to the Friedmann equation (I.2.9) and look at the right-hand side, you see that the vacuum energy is a constant positive part of the density ρ. However this constant is multiplied by a^2, and thus this term (if present) will under all circumstances grow faster with respect to the second term that corresponds to the curvature constant k. What this means is that a universe that goes through such an exponential phase will blow up and effectively become flat. The situation is somewhat analogous to the claim by some Dutch people that their country is flat; it is indeed effectively flat, but not really. It is better to say that the curvature radius of the earth is much larger than their visual horizon. Going back to the universe, what this means is quite interesting. If you could turn the vacuum energy on for a limited amount of time, the exponential expansion would basically flatten out the universe. This is a vital observation because it would furnish a dynamical mechanism by which the universe drives itself to that unique point in the solution space where k is effectively zero! The universe would end up being flat, becoming open *independent* of the initial situation.

What do the experiments tell us? The data unequivocally suggest that there must have been a brief period in the very early universe, where it expanded exponentially. And that brief period of *cosmic inflation* as it is called is the reason we find ourselves in a flat ($k = 0$) universe now. We have pointed out two serious problems where the standard cosmological model clashes with the data, and both are resolved in the inflationary scenario to which we turn next.

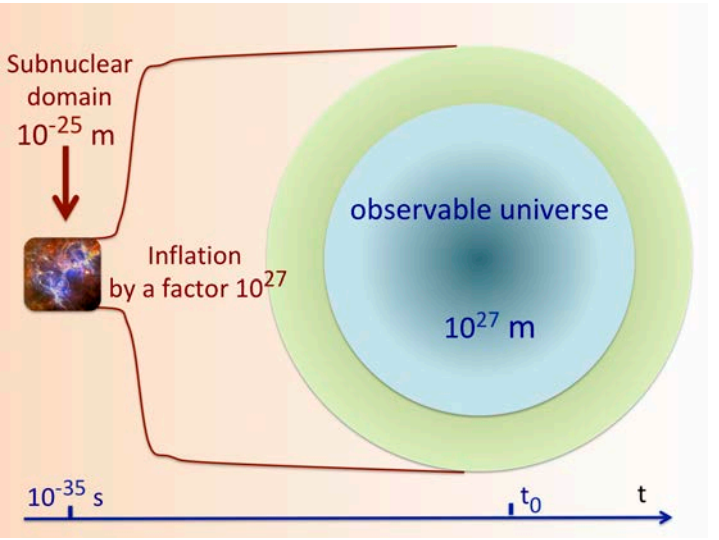


Figure I.2.19: *Cosmic inflation.* Inflation makes the present observable universe fit in the expanded image of a causal domain of subnuclear size.

The inflationary scenario.

The inflationary scenario involves a non-vanishing vacuum expectation value of a so-called *inflaton field*, presumably some scalar field that has not really been identified. In a very, very early stage of the universe, say, at $t \simeq 10^{-35}$ seconds, due to the cooling of the universe this inflaton field gets stuck in a metastable vacuum state. This means that it generates a constant vacuum energy in the universe, and this will last for about $t \simeq 10^{-32}$ seconds, after which it will decay to a new lower zero energy ground state. During this period with the non-vanishing vacuum energy present, the universe would inflate the linear dimension of the uni-

verse by a factor of 10^{27} (corresponding to 10^{81} for the volume). Inflating a causal domain by such a huge factor solves the horizon problem as is indicated in Figure I.2.19. The epoch ends with a phase transition of the early universe as a whole. The latent heat released in this transition will be converted into ordinary matter. Such a scenario implies a drastic revision of the very early stages of the standard cosmological model. Note that though the time periods appear to be extremely short, this is only relative, the inflationary epoch lasts a 1000 times the age of the universe at that time! You could therefore equally well say that it took 'ages.'

There is one other observational aspect of early universe cosmology that this scenario gives an answer to. The enormous inflation factor basically implies that our whole observable universe originates from an extremely small domain before inflation started. The domain would be so small that the physics within that domain would be governed by quantum theory. That particularly implies that within such a domain of size Δx there are substantial quantum fluctuations, and that these fluctuations have a flat, scale invariant spectrum, meaning that their amplitude is independent of their wavelength. These small wavelength fluctuations ($\lambda \leq \Delta x$) are blown up to large-scale inhomogeneities by the inflation. And it is believed that these inhomogeneities are the seeds of large-scale structures in the subsequent evolution of the universe. Knowing the initial spectrum at the end of the inflationary epoch allows one to predict what the inhomogeneities and anisotropies in today's cosmic background radiation would be. And indeed the scale invariant initial spectrum evolves in a highly non-uniform distribution with damped oscillations which agrees extremely well with what has been observed by space observatories like WMAP and PLANCK as is shown in Figure I.2.20.

This surprising scenario combines knowledge from the microscopic realms of quantum field theory, with knowledge from general relativity and cosmology and allows for a so-

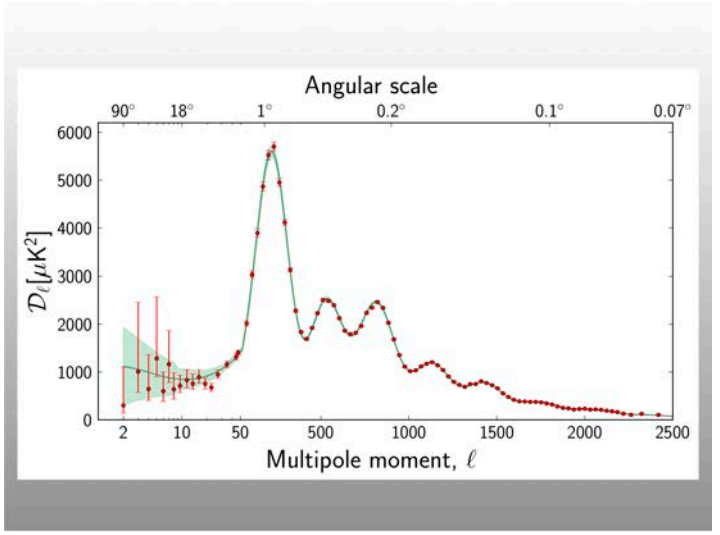


Figure I.2.20: *CMB anisotropy.* The inflation blows tiny quantum fluctuations in the initial causal domain, up to large-scale inhomogeneities that are believed to be the seeds of larg- scale structure in the present universe. These show up in the angular correlations in the spectrum of temperature anisotropies of the cosmic background radiation. From this data the three energy parameters and the cosmic curvature constant in the model can be determined. (Source: PLANCK mission)

lution of both the horizon and the flatness problem of standard cosmology. Scenarios of this type were proposed and developed in the early 1980s by Alan Guth from MIT, Andrei Linde presently at Stanford University, and Paul Steinhardt presently at Princeton University.

Splendid observations. Having presented these fascinating theoretical considerations, let us briefly review the stunning progress that has been made in observational astronomy and cosmology. The fundamental observational parameters in the cosmological models are the energy densities Ω_i , and the Hubble constant and these basically tell you what the curvature constant k is. Two completely different techniques have been used:

(i) The measurement of very distant Supernovae type I

events. These are basically very remote sources that allow us to extend the Hubble law plot of Figure I.2.11. A great experimental effort by Saul Perlmutter and collaborators (1998) managed to expand the diagram by a factor of ten, and the spectacular discovery they made was that the plot does no longer stay linear but is curving upward. This means that at large distances we see the expansion accelerate. They fitted the data and extracted the Ω values, and clearly obtained a positive contribution for the vacuum term. With Adam J. Riess and Brian P. Schmidt, Saul Perlmutter was a co-recipient of the Nobel prize for Physics in 2011, and the prize was awarded 'for the discovery of the accelerating expansion of the Universe through observations of distant supernovae.'

(ii) The measurement of the curvature through measuring the anisotropy in the Mmcrowave background radiation also gives – among many other things – the Ω values. There have been a number of space telescopes to do this kind of work: first the COBE (1992), then WMAP (2003) and most recently the PLANCK (2013) mission, with again startling results. For this line of research the Nobel Prize in Physics of 2006 was awarded jointly to John C. Mather and George F. Smoot of the COBE collaboration 'for their discovery of the blackbody form and anisotropy of the cosmic microwave background radiation.'

Concerning the relative energy densities, the upshot of these experiments is summarized in the energy piechart depicted in Figure I.2.21. After the PLANCK mission the preferred fractions are: 68.3 % is in the form of vacuum or dark energy, 26.8 % in the form of dark (not luminous) matter and only 4.9 % is in the form of ordinary luminous matter. The conclusion is crystal clear: we are living in a vacuum dominated, flat universe! What that means ultimately is also not hard to understand. The remarkable message is that 95 % of all the energy in the universe resides in the dark matter and energy components, and is therefore in a form that is unknown to us! It reminds us of the words of the Chinese philosopher Lao Tzu: 'The more

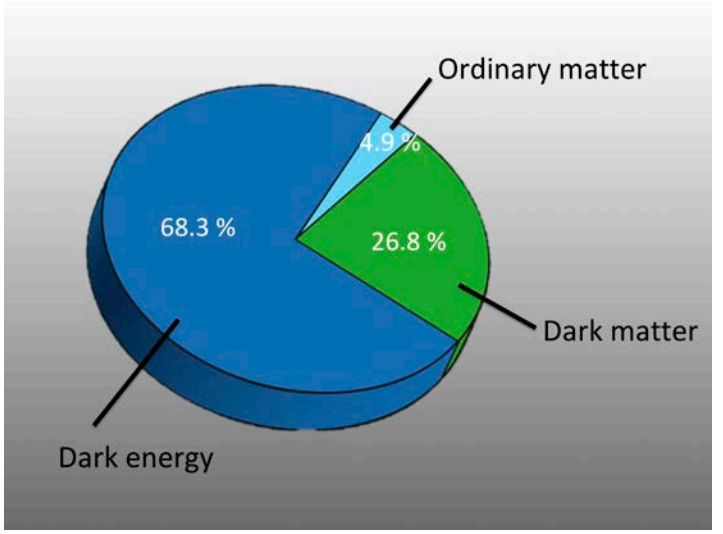


Figure I.2.21: *The energy piechart.* A piechart of the relative contributions of dark (vacuum) energy, dark matter and ordinary matter as determined by the WMAP and PLANCK space observatories. Conclusion: our universe is vacuum dominated.

we know, the less we understand!' More than anything else, science is the story of work in progress, every time reminding us of our ignorance, and forcing us to cope with it. Or to find creative ways to beat it. Indeed, science will always run into new walls, or to be more encouraging: new profound challenges.

Today's challenges. We have to conclude that looking at the presently available data, this consistent and convincing, evidence-based inflationary model of the evolution of the universe still leaves us with some big puzzles.

Dark matter. The first is the question what actually is dark matter. Clearly this is a question that has been taken up by the particle physicists who built their Large Hadron Collider (LHC) at the European accelerator center CERN in Geneva. They are presently hunting for a new particle type that would fit the profile of dark matter. Such particles should be 'sterile,' meaning that they interact very weakly with ordinary matter and they should be massive in order

to cause the gravitational effects we observe. They are expected to form a species of so-called WIMPs: Weakly Interacting Massive Particles. Theoretical candidates are for example the species of lightest supersymmetric particles (a necessary ingredient of String Theory), or various types of massive particles that are called 'sterile' neutrinos (fitting in certain Grand Unified Theories).

Dark energy. The second even more profound puzzle is the observed non-zero value of the cosmological constant, or vacuum energy. It is 'small' but definitely non-zero, and the question is wheter we can find a theoretical explanation for its existence and its magnitude. The irony is that physicists have for quite some time been looking for arguments why it would have to be strictly zero exactly because there was an extremely strong bound on it from observation. They looked for a principle that would protect the zero value of the cosmological constant, like the gauge principle protects the zero mass property of the photon. Needless to say that they didn't succeed, fortunately in fact, because now we know that it is not zero to start off with. Answering this question requires a fundamental insight into the nature of the vacuum, and so far there is no way to calculate the quantum energy of the vacuum from first principles. Such an explanation should also allow us to make a first estimate of its magnitude, because in spite of the fact that it is the dominating energy content of the universe, its actual value is mesmerizingly small: $\Lambda = 1.1 \times 10^{-52}\,\mathrm{m}^{-2}$. This mass energy density is about four protons per cubic meter, which amounts to $\rho_{\mathrm{vacuum}} = 5.9 \times 10^{-27}\ \mathrm{kg/m}^3$.

From a theoretical point of view, the conclusion is that the De Sitter solution, which was discarded for a long time as physically irrelevant, has made a glamorous comeback, and presently plays a vital role in understanding the deep past, as well as the present and future of our universe. Remarkable!

Figure I.2.22: *Magritte: the pilgrim (1966).* My title for this intriguing surrealistic painting would be: 'Let's face the void, and void the face.' If that isn't a deep thought, then neither is its negation! (Source: ©'Photothèque Magritte / Adagp Images, Paris)

Much ado about nothing. The handicap of generalists is that they know virtually nothing about almost everything, and the handicap of nerds is that they know virtually everything about almost nothing. What? Knowing everything about nothing? I wish it were true. Closer inspection shows that the science in-crowd knows little to nothing about nothing. Scientists remain silent, but spend sleep-

less nights worrying about nothingness. Imagine some DOE Innovation Initiative inspection team performing a lab raid and asking what you are doing with all that taxpayers' money, and you would have to answer that you are working on 'nothing.' Oh yes, you are just mucking around, are you? That would undoubtedly result in you taking a deep dive in the cool lake of depression. Career-wise I would avoid talking about the void.

You would think that empty space – the vacuum, the void, nothingness – is a trivial no-brainer not worth pondering about. Note however, the following important remark that the legendary physicist John Archibald Wheeler made at some point: 'No point is more central than this, that empty space is not empty. It is the seat of the most violent physics.' Here, a deep truth appears to be lurking. A quantum truth.

The reason we don't need to talk about it in physics is because in real experiments we are always dealing with energy differences. We compare energies and the energy of the vacuum 'drops out.' We therefore can set it equal to zero if we would like to do so. This is fortunate, because we do not know how to calculate the vacuum energy from first principles, and all 'serious' efforts to do so typically give infinity as an answer. This means that the void is challenging our deepest scientific intuitions. General relativity is a comrade in arms, because as we saw, it *is* sensitive to something that other theories would not detect. Space-time itself allows for an *absolute* measurement of the energy including that of the vacuum. And moreover, space-time measurements have just told us that the vacuum energy is not just non-zero, it is in fact the dominant form of energy in our universe!

This much is certain, 'nothing' does not exist and the notion of nothingness is an apparent delusion. What does exist is our ignorance about it.

So, what's so tricky about nothing? An average fish would reply: 'Well, no fellow-fish, no water-plants, no play-rock and no gravel on the bottom.' But what the average fish would never say is: 'no water.' The fact that nothing is something in which he couldn't exist doesn't enter his fishy head. The average person by now understands damn well that without air he is going to choke, but apparently in nineteenth century educational institutions, that simple fact still had to be demonstrated by putting a little bird under a glass bell and pumping out the air. Just to prove that 'nothing' can also be quite harmful. Causing all sorts of panic because of the 'unbearable heaviness of not-being.' 'To be or not to be,' remains the question. Having answered that, 'to understand or not to understand' is the next question. □

The physics of geometry

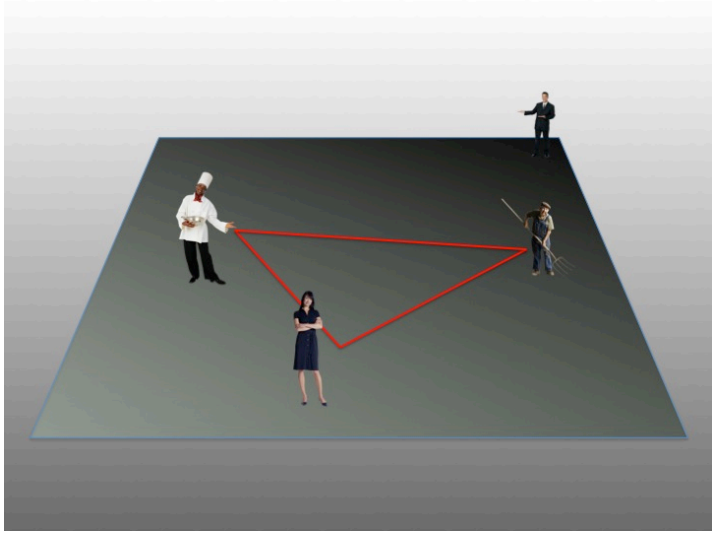
With the advent of the theories of relativity and gauge theories for the description of the fundamental forces, a new golden age for geometry in the realm of physics emerged. This section on 'the physics of geometry' will give you an introduction to the basic notions of geometry that have played a crucial role in modern physics.

We will talk about the notion of curved spaces (smooth manifolds*) and which concepts are essential if one wants to do physics on and with them. Some aspects of topology are mentioned like* homotopy*, because it leads to an alternative way of understanding why certain physical quantities turn out to be quantized and conserved.*

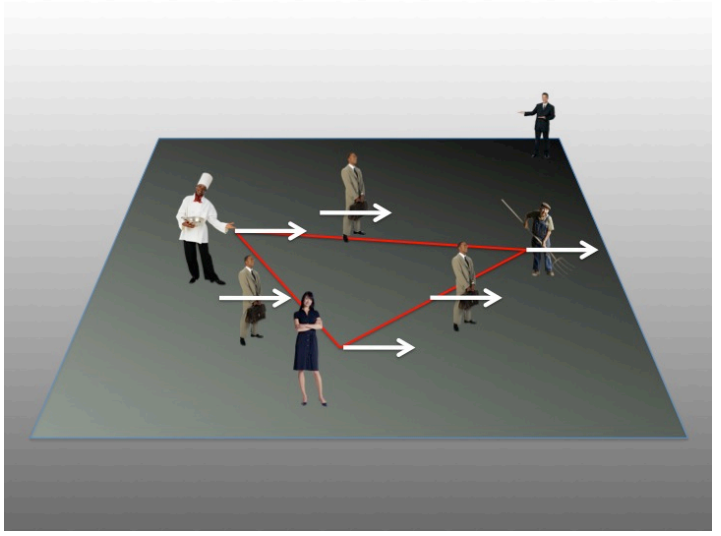
Introducing regular coordinates on curved spaces often forces us to define overlapping patches or charts. One may introduce vectors that live in the tangent space *of some point, whereas the collection of the base manifold and all its tangent spaces form the so-called* tangent bundle *of the space. Now we may study the transport of vectors along paths. This leads to the notion of* holonomy *which is linked to the integrated curvature over a region of space. We complete this lightning review by summarizing the relations between the* metric*, which provides a local sense of distance and size, a* connection *which connects vectors in different points, and a local* curvature *which is very much like the field strength in a gauge theory.*

After this pure geometrical part we show the close relationship of gauge theories, describing the non-gravitational interactions, with fiber bundles*. This provides a geometrical understanding of the gauge potential or connection* A_μ *and a gauge invariant field strength or gauge curvature* $\mathrm{F}_{\mu\nu}$ *. This representation of gauge theories opens the door for understanding their topological features, like the existence of magnetic fluxes or monopoles, and more generally to the notion of topological charges and quantum numbers. This section aims to highlight the intimate relationship between physics, geometry and quantum theory.*

Living in flatland. When you talk about a space, most people have the natural inclination to think of a flat Euclidean space, like the plane denoted by $\mathbb{R}^2$ **.** And on a plane life is simple not only for the Danes and the Dutch, but also for physicists. It is simple to choose a coordinate grid to label the points in the space. The shortest distance between points are straight lines, and to define vectors like momenta and the forces of electric fields is also simple. You just draw arrows based at a point in the space because a flat Euclidean space is also a vector space. There is no distinction; you may think of vectors as living in the 'same' space as you. On flat Euclidean space we can define functions and derivatives of functions (basically vectors), as well as their integrals. If we have a particle moving on the plane in a potential $\mathrm{U}(\mathbf{x})$ then it will experience a force $\mathbf{F}(\mathbf{x}) = -\boldsymbol{\nabla}\mathrm{U}(\mathbf{x})$ which corresponds to a field of vectors over the plane. In other words we can do calculus,



(a) Triangle in flat space, sum of angles is 180^{o} .



(b) Parallel transport of a vector in the plane.

Figure I.2.23: *Carrying vectors around.* Parallel transporting around a closed loop in a flat space has no net effect on the orientation of the vector.

and therefore flat space is the basic example of a space or manifold that is *differentiable*.

Parallel transport of vectors. In Figure I.2.23 we show a Master Chef and two of his branch managers running annexes in other parts of town. He wants to send a secret recipe around in the form of a vector, and it is in the orientation of the vector in which the subtle balance of spices that earned the Chef his Michelin star is encoded. So, it is crucially important to preserve the direction the direction in which the vector points, implying that the Chef cannot send the recipe by mail. He decides to hire a messenger, an apprentice so to say, in a grey suit and with a leather briefcase to carry around the vector. The messenger should take care to *parallel transport* the vector. This is not hard: while moving along the shortest route consisting of straight line segments, he has to ensure that the angle between the vector and the direction of his motion (the path) stays the same. The Chef has ordered him to pass by again at the end of the trip, so he can check whether he did the parallel transporting correctly. And as you see the apprentice succeeded in perfectly performing his task, as is confirmed by the independent juror standing in the corner.

In this subsection we have mentioned some features of flat space that are so natural that you wouldn't think of them as particularly interesting. However, what we will explore in the remainder of this section is that in a curved space, these simple concepts will become much more involved

Curved spaces (manifolds) and topology

Modern physics makes use of the mathematical knowledge about curved spaces or manifolds, both in the theory of relativity, but also in the theories that describe gauge interaction between elementary particles. What we want to introduce are what is known as differentiable manifolds, curved spaces that look locally like flat Euclidean space and therefore globally allow for defining functions and their derivatives (and vectors). These are spaces on which one can consistently define calculus, a necessary tool to describe dynamical systems in such spaces. And that is what physicists like to do. We start by defining some elementary notions of *topology* , and then add the ingredients of differential geometry like coordinate systems, vectors, metric and curvature.

Positive and negative curvatures. It is easy to imagnine curvatures of a surface when we embed it in a higher dimensional Euclidean space. The surface can be defined by an algebraic equation.

Consider for example spheres S^n, these are defined by an equation $x^2 + y^2 + \ldots = 1$ in $(n+1)$-dimensional Euclidean space $\mathbb{E}^{n+1}$. In Figure I.2.24 we show the spheres S^0 (two points), S^1 and S^2.These spaces are finite or compact. They can be obtained from one another by suitable rotation in the Euclidean space two dimensions higher.

Spaces of negative curvature are for example hyperbolic spaces. In Figure I.2.25 we depicted two hyperbolic surfaces H^2 and the corresponding equations to contrast them with the two-sphere. One of the hyperboloids consists of two disconnected sheets while the other has only one sheet. These spaces are infinite. These two spaces can be generated by rotating a given hyperbola in the plane about an appropriate axis in that plane. The sphere is by definition the set of all points that have a fixed Euclidean distance to the origin. The double-sheeted hyperboloid can be thought of as the set of all events in three--dimensional Minkowski space, at a fixed space-time interval from the origin. The sheets also represent the three-dimensional energy momentum vectors $(E, \mathbf{p})$ of a particle with finite rest mass m satisfying $E^2 = m^2 + \mathbf{p}^2$.

Topological features. A topological feature or characteristic is one that doesn't change under a *continuous de-*

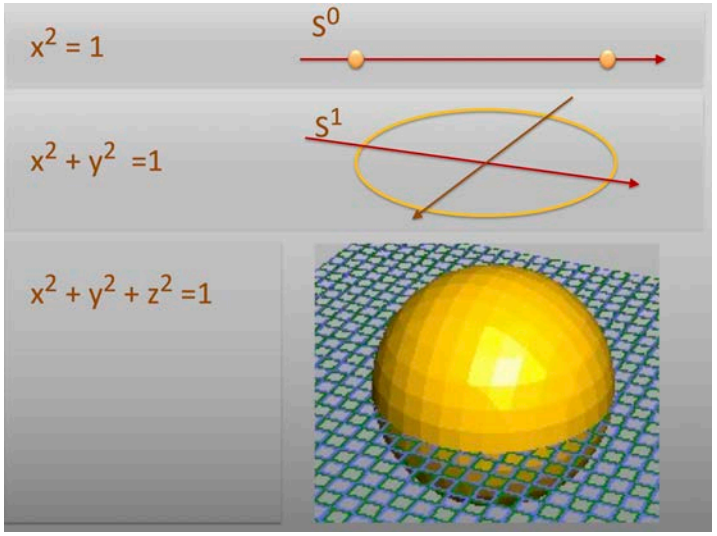


Figure I.2.24: *Three spheres.* This figure shows the spheres and their embedding equations. The 'zero'-sphere S^0, described by the equation $x^2 = 1$, consists of two points. The circle or one-sphere S^1 is defined by the equation $x^2 + y^2 = 1$, and S^2 by its natural higher-dimensional extension.

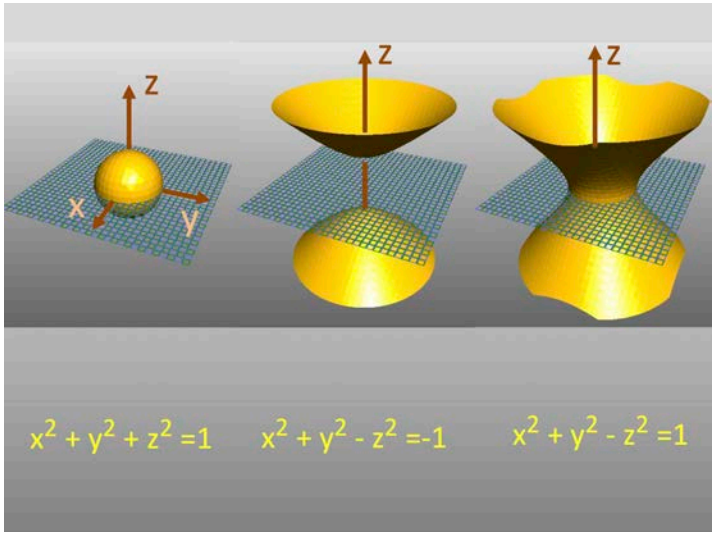


Figure I.2.25: *Hyperbolic planes.* This figure shows two different two-dimensional hyperbolic spaces and their embedding equations. They are closely related to the equation for the two-sphere, and differ by additional minus signs. The sphere has positive curvature, while the hyperboloids have negative curvatures.

formation of the space. Cutting and pasting the space is not allowed. Topology is like rubbersheet geometry where stretching and shrinking in any direction is allowed but tearing the sheet is not. Two spaces are topologically equivalent if they can be continuously transformed into each other.

Boundaries and holes. Let us start with one-dimensional spaces like smooth curves. For example a line segment is smooth and has two point-like boundaries. But we could also consider a closed curve. It may look like a circle or the number zero which has no boundary, but it has a hole. The figure 'eight' is also closed (has no boundary) and it has two holes but now it is not a one-dimensional space because it has a singular point where the lines split and where it therefore is locally *not* like $\mathbb{R}^1$.

Connectivity. The spaces just mentioned are all *path-wise connected*, meaning that for any two points one chooses there is a path connecting them. The number 'I0' as a space is not connected it has two disconnected components: the 'I' and the '0'. One open component 'I' has two point-like boundaries, and the closed component has no boundary. If we consider any two points on the line segment then these can be connected by some path, and all paths can be continuously deformed into each other. Taking two points on the '0' or a circle, we find that there are many possible paths that connect the two points. These paths may wind an arbitrary number of times around the hole and such paths cannot all be continuously deformed into each other. We say that the 'I' is *simply connected* whereas the figure '0' is *multiply connected* because there are topologically distinct paths. So we are invited to further refine the notion of connectivity. Let us do that after we have moved the discussion one dimension up and consider smooth two-dimensional surfaces.

The simplest finite, two-dimensional spaces have the topology of a disc. It is simply connected and it has one boundary with the topology of a circle. Note that a boundary in two dimensions in general is a disconnected union of one-dimensional closed curves, which are topologically speaking circles. If the boundary has more than one component, the space becomes multiply connected.

To imagine a curved space one may for instance think of the two-dimensional surface of a sphere or torus as embedded in a three-dimensional flat Euclidean space $\mathbb{R}^3$. If you look at a small neighborhood of any point in these curved spaces S^2 or T^2, you see that locally, it is everywhere like the flat space $\mathbb{R}^2$.

It is only after you enlarge your horizon that it becomes clear that the sphere and the torus are quite different globally from flat space and from each other. Indeed the study of curved spaces descended on us with the insight that the earth turned out to be not flat. Both are globally *compact* meaning that they are finite: it takes a finite amount of paint to cover the two-sphere for example, whereas flat space is infinite and non-compact. Similarly a three-dimensional sphere would have a finite volume.

Spheres and tori are finite spaces, and they also have the property that they have no boundary. Indeed curved spaces can be finite and not have a boundary. Yet, they do have a different *topology*; for example the two-torus has a hole in it while the two-sphere has not. This means that the connectivity properties will differ and this in turn implies that the physics in the one space may exhibit features different from the other.

For two-dimensional manifolds without boundaries (also called closed Riemann surfaces) the number of holes is the only topological invariant characterizing the manifold. A *pretzel* is therefore distinct from a *donut*, not only qua substance and taste but also topologically, as it has two holes.

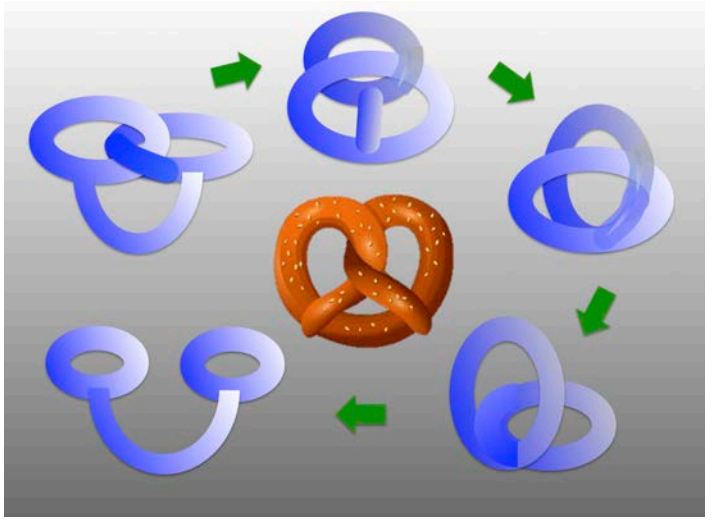


Figure I.2.26: *The pretzel-transformation.* This figure shows in clockwise steps how a self-linked pretzel (top left) can be smoothly unlinked (left bottom). It is a nice example of a topological deformation as prsented by Martin Gardner in his book (1987) on mathematical recreations.

In spite of the pretzel's simplicity its topology is surprisingly counter-intuitive as we have illustrated in Figure I.2.26. One can clearly imagine the left and right parts of the pretzel to be interlinked like the real pretzel in the center and schematically depicted at the top on the left. It appears like yet another two-dimensional closed surface which is topologically distinct with some two holes and a half! Is it really? The answer is: No! There is a well-known smooth topological deformation that corresponds to a smooth unlinking of the pretzel. In the figure we depicted the subsequent steps in the so-called pretzel-transformation which shows how a self-linked pretzel can smoothly be unlinked. I always imagined that this somehow must be of use if you end up in the unfortunate situation of being handcuffed for some reason, for example because of stealing pretzels!

Homotopy. An important topological characteristic is denoted as the *connectivity* of a space, which can be probed

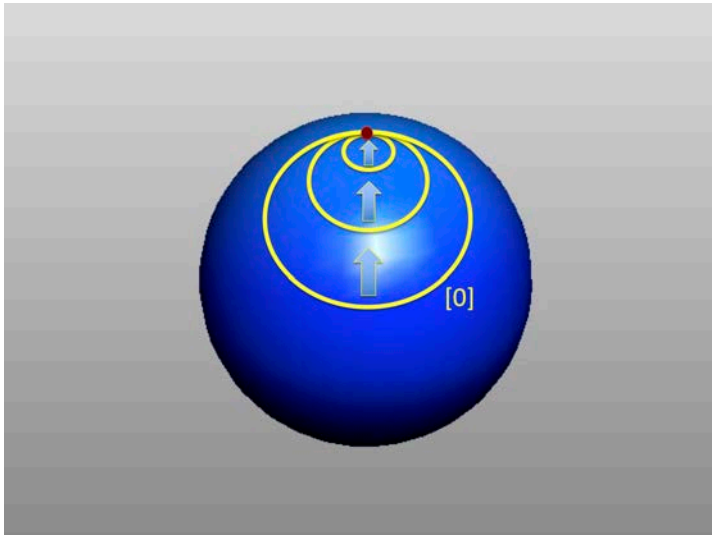


Figure I.2.27: *The two-sphere is simply connected.* If we take a point and consider closed loops starting and ending in that point, all possible loops can be smoothly deformed into each other, and all loops are contractable to the point. There is only one trivial homotopy class, denoted by $[0]$.

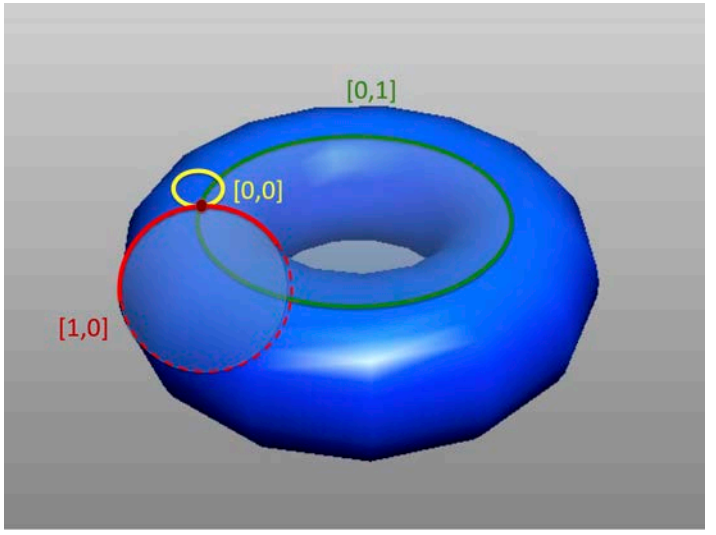


Figure I.2.28: *The two-torus is multiply connected.* We have depicted three loops through a point. The yellow one is contractable and belongs to the trivial class $[0,0]$. The green (red) one winds once around the large (small) hole and is non-contractable, and it belongs to the class $[0,1]$ ($[1,0]$).

by studying maps of closed paths or loops into that space. The loops that can be continuously deformed into each other are called *homotopic*. Homotopy is an equivalence relation. Two loops are either homotopic or not. Having such a relation allows you to divide the space of all maps of loops into distinct classes, *homotopy classes* in this case. For example if you draw a closed loop on a sphere, this loop can always be smoothly contracted to a point. The popular wording of this fact is that 'You cannot lasso a basketball.' From Figure I.2.27, we see that all loops on the sphere can indeed be deformed into each other and can smoothly be contracted to a point, so there is only one *homotopy class*, the trivial class denoted by $[0]$.

However, if you look at closed curves on a torus, then there are many possibilities. There are loops that can be simply contracted to a point, then there are loops that wind around the big hole, like the big circle on the outside, or closed curves that wind around that hole an arbitrary number of times. There are also loops 'perpendicular' to the previous ones, going around the small hole a certain number of times. We have illustrated the situation for the torus in Figure I.2.28, where we have drawn three examples. The yellow loop is contractable and therefore belongs to the trivial class which we denote by $[0,0]$. The green loop encircles the large hole once: it is non-contractable and belongs to the class $[0,1]$. The red loop encircles the small hole once, and cannot be deformed to either of the other two, since it belongs to another class $[1,0]$. In general a loop that winds m times around the small hole, and n times around the big hole belongs to the class $[\mathrm{m},\mathrm{n}]$. Think for example of a hiking boot as a closed two-dimensional surface: it may have ten holes for the shoe lace to go through. When I have tied the knot I should have created a closed loop belonging to a non-trivial class.

Having defined and labeled these classes in a systematic way we can go one step further and ask if additional prop-

erties can be assigned to them. The first thing that comes to mind is: can we assign an orientation or direction to them. This can be done by putting an arrow on them, and this allows you to assign negative winding numbers.

The first homotopy group. A nice property of closed paths is that we can compose them by connecting the end point of the first loop to the beginning of the second loop and so create a new closed path. This composition rule induces a map, or more precisely a multiplication rule for the homotopy classes: $[...]_1 \odot [...]_2 = [...]_3$. So here we have a set of objects (a set whose elements are classes) that we know how to multiply but there is not such a thing as addition defined. This means that this set forms a *discrete group* where the unit element corresponds to the trivial class of contractable loops, while the inverse element is the class corresponding to the opposite winding number. This group is called the *first homotopy group*, or *fundamental group* and can be determined for any manifold.

A question that may come to mind is: What does this have to do with physics? The answer is: quite a bit. In fact we have already seen examples of it. The notion comes up if you want to discuss line integrals of some field along a closed curves, as we did for example with the vector potential. The loop integral corresponded to the enclosed magnetic field, or the magnetic flux going through the loop. The group structure tells you how these magnetic fluxes 'add.' And as it turns out these fluxes can have highly unexpected composition rules once one studies phases, not of electrodynamics, but of non-abelian gauge theories. These considerations have also important applications in the study of quantum interference effects and the quantum statistics of particles. These are topics we will get to in later chapters of the book.

Higher homotopy groups. In higher dimensions there are more possibilities to consider. For one you may think of higher dimensional holes that correspond to non-contractable maps of higher dimensional spheres into the manifold, and these in turn form higher dimensional homotopy classes. So the second homotopy group tells you how many topologically inequivalent ways there are to map a two-sphere (a closed two-dimensional surface) into the manifold and how those maps can be composed. Finally, the zero-dimensional homotopy classes label the disconnected components the space under consideration.

Coordinate systems. You may wonder why it took so long for mankind to figure out that the Earth's surface we are living on is a space that is not flat but curved. The reason is that on a local scale the world is basically flat and our naive expectations work well as long as you stay nearby. So, if we live in a curved space it has to be a space that is locally like Euclidean space. A space that is everywhere 'locally flat' is a space that we call smooth because we can systematically extend the whole mathematical apparatus of calculus concerning differentiation and integration of functions which we originally defined on flat space. So we may expect to be able to give adapted mathematical descriptions of the physical laws if we move from flat to curved spaces, as relativity tells us to do.

Euclidean space and coordinates. The first question that arises is to choose coordinates on the space. The choice of coordinates are often naturally suggested by the symmetries of the space. You could think that the symmetries generate the whole space from a single point, an 'origin.' For example, flat space has *translation symmetry*, we can move from any point to any other by performing a translation. $\mathbb{R}^n$ has n independent orthogonal directions in which a given point can be moved, and the natural choice for coordinates is therefore the Euclidean coordinate system $\{x_1, x_2, \ldots, x_n\}$.

Curvilinear coordinates. But nobody forces us to use that coordinate system. In fact as soon as we start considering a particular setting in that space, for example, we may single out a particular point as the center of our space. Think of how we used to put the Earth in the center, and

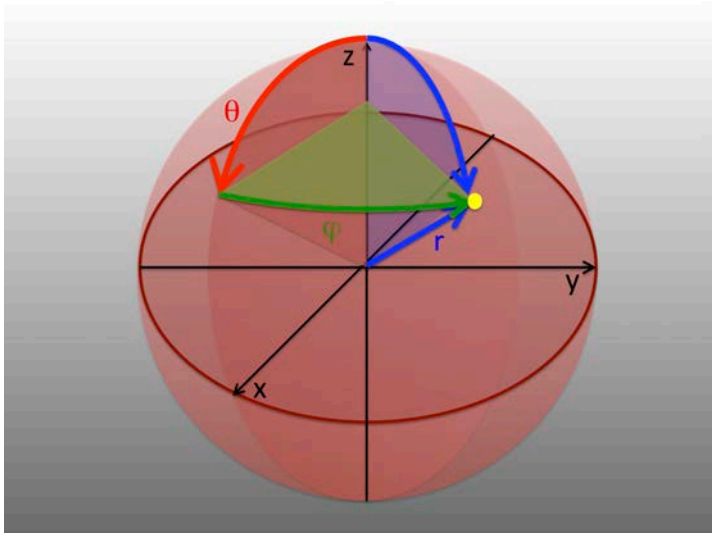


Figure I.2.29: *Spherical coordinates.* The definition of spherical coordinates in $\mathbb{R}^3$, denoted by r, θ, and φ .

after the Copernican revolution we put the Sun in the center and so on. Fixing one point as special, we break the translational symmetry, but still have the *rotational symmetry* which leaves that point fixed. And that symmetry naturally suggests another choice of coordinates, the system of spherical coordinates $\{\mathrm{r}, \theta, \varphi\}$, as depicted in Figure I.2.29 . It is as if we think of $\mathbb{R}^3$, as built up of a point plus a continuous stack of concentric spheres around it. Similarly we may want to use cylindrical symmetry with coordinates $\{\rho, z, \varphi\}$, where we think of the space built up from a line with a continuous stack of concentric cylinders around it. And we have already seen that in many physical applications such orthogonal curvilinear coordinate systems are much more convenient; they lead to a convenient framing of the problem that makes it easier to obtain solutions. For example, if I have a current through a straight wire along the $z-$axis like in Figure I.1.18, the problem becomes cylindrically symmetric, and the magnetic field $\mathbf{B}(\mathbf{x})$ will have only a φ-component that will only depend on the radial coordinate: $\mathbf{B} = \mathrm{B}_\varphi(\rho)\,\hat{\mathbf{e}}^\varphi$.

Spherical coordinates. The observation that we think of spaces generated by symmetries is useful if one wants to think of typical curved spaces which exhibit those symmetries. Indeed if we think of the three-dimensional rotations just mentioned, and we take an arbitrary point in $\mathbb{R}^3$, that point will indeed trace out a two-sphere, which is a highly symmetric two-dimensional space. So if we 'throw away' the radial coordinate, we are left with an orthogonal coordinate system on the sphere consisting of the two angles, the *polar angle* θ , running from 0 to π and the *azimuthal angle* φ running from 0 to 2π, as we have been using all along. Do these coordinates cover the sphere well? Not really, it turns out.

Coordinate singularities. The north and the south pole are clearly problematic. In these points the coordinate system breaks down, whereas the θ coordinate is well defined, the φ angle is not. There is no sensible way to assign a definite φ angle to the poles. Note that the real geometry of the sphere is completely smooth at those points. The poles are regular points just like any other point on the sphere. The problem is not the space, but the coordinate system we have chosen. To solve this coordinate problem in general one first has to accept that it is not possible to choose a single coordinate system that covers the whole sphere without singularity. There is a topological obstruction to doing that following on from the *hairy ball theorem*. This theorem states the easy to imagine fact that it is not possible to comb a hairy sphere. Just try doing it and you will quickly find out that there is always a point in which the hairs meet in opposite directions. This means that there is no globally defined, non-zero tangent vector field, or conversely, that any vector field on a sphere has to vanish at least in one point. And that fact implies that we cannot have a single globally defined coordinate frame of orthogonal unit vectors on the two-sphere.

Patches or charts. Knowing this fact, the best we can do is to cover the sphere by defining two coordinate *patches* or *charts*, that together cover the sphere and have an overlap so that we can identify points on the two maps that

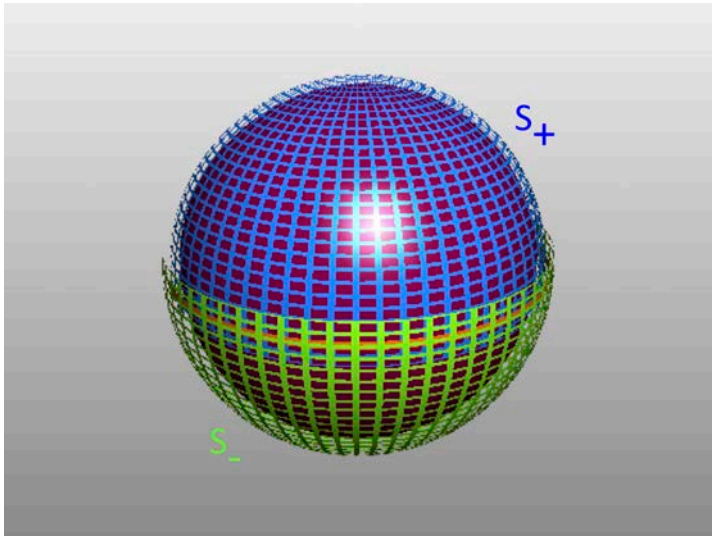


Figure I.2.30: *Two coordinate patches.* The two-sphere is covered by two overlapping coordinate patches $S_\pm$ to avoid coordinate singularities. There is a *transition map* which identifies points (and tangent vectors) in the overlap region. In this case we have $\varphi_+ = \varphi$ and $\varphi_- = -\varphi$. and $\theta_- = \theta_+ = \theta$.

correspond to the same points on the sphere. For example, one may define one patch covering a little more than northern hemisphere and call it S_+ , and the other a little more than the southern hemisphere and call it S_- , as indicated in Figure I.2.30. The overlap between the patches is then a narrow band with the equator in the middle. Each patch has the topology of a disc and so we can put a regular coordinate grid on it. Now we can define a *transition map* on the overlap, which provides a map between the coordinate systems in both patches. And this map should be smooth as well. At this point we have succeeded in making an *atlas* of the sphere consisting of two *charts*, each of which can be smoothly mapped onto a flat page by a stereographic projection, which you may have encountered in high school geography classes. This is the way to deal with the complications that arise in defining coordinates on a sphere, and this allows us to globally define smooth functions and their derivatives, to define paths and vectors, all the things physicists and mathematicians need and love. With this construction we have shown that the two-sphere also is a *differentiable manifold*, a curved space where we can do calculus. A differentiable manifold is a space that is locally like Euclidean space, and globally looks like a smooth patchwork of pieces that are much like flat space, sewn together in a consistent way by a network of smooth transition (sewing or gluing) functions.

Distance and path length. So far we have talked about topological characteristics of manifolds but that leaves the important aspect of form and scale undetermined. How long do I have to walk to get from A to B, that's the question! Mathematicians like to say that to settle it we have to add more structure to the space. In order to introduce the concept of size or distance we have to define a *metric* on the space. In flat space we know that the shortest distance between two points corresponds to the straight line between them. And the distance is calculated by applying the Pythagorean theorem. If we consider any smooth path in flat space, we can calculate its length by successively applying the theorem to infinitesimal segments of the path and adding the results. If the points are nearby, we have for the distance ds , that $ds^2 = dx^2 + dy^2$. If we start by defining a smooth *path* as a one-parameter curve $\gamma(t) = \{x_\mu(t)\}$, the *tangent vector* to the curve at the point $x(t_0) = \gamma(t_0)$ is just like the 'velocity' vector $\mathbf{v}(t) = d\mathbf{x}/dt\big|_{t_0}$. The length L_{ab} of the curve between two points $\gamma(a)$ and $\gamma(b)$ is now quite naturally defined as the integral:

$$L_{ab} \equiv \int_a^b |\mathbf{v}(t)| dt \,. \tag{I.2.15}$$

in a different, more familiar wording, the distance traveled is just the magnitude of the velocity integrated along the path over the appropriate time interval. It is this notion of path length that we like to generalize to curved spaces.

Metric and line element. To calculate the path length in a curved space we need a local definition for the infinitesimal distance ds which specifies the local (x-dependent) definition of an infinitesimal length. Once we have chosen

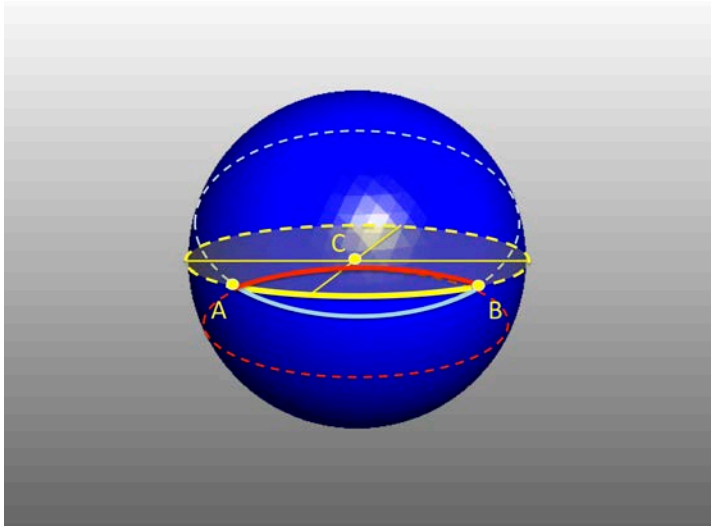


Figure I.2.31: *Shortest distance.* The shortest path between two points A and B on the sphere is given by the segment of the yellow 'great circle' which is defined as the intersection of the plane through A , B , and the center of the sphere C, and the surface of the sphere. The blue and red circles are smaller and yet yield longer paths between the points A and B.

coordinates $\{x_i\}$ on the space, where i runs from 1 to d, the dimension of space, then we formally define the so-called *line element* ds as:

$$ds^2 = g_{ij}(\mathbf{x})\,dx^i\,dx^j \tag{I.2.16}$$

where the symmetric matrix $g_{ij}(\mathbf{x})$ is the so-called *metric tensor*. In flat space we saw that $ds^2 = dx^2 + dy^2$ and the metric thus corresponds to the unit matrix everywhere. If we choose polar coordinates r and φ, the line element would be $ds^2 = dr^2 + r^2 d\varphi^2$, and the metric a diagonal matrix $(1, r^2)$.

If we put a symmetric mass distribution around the origin, then the space would be curved as we illustrated in Figure I.2.4, where the two-dimensional surface embedded in $\mathbb{R}^3$ would be defined by fixing $z = f(r)$ with a function f interpolating between some constant $f(0) = -a$ and $f(r \to \infty) = 0$. The radial length measured along the surface will now change, and indeed the metric would change in that $g_{rr} = 1 + (df/dr)^2$. The metric on the two-dimensional surface is *induced* from the trivial metric on $\mathbb{R}^3$ by substituting $dz = (df/dr)\,dr$.

It is important to realize that in principle there are many possible choices for the metric on a manifold, the only restriction being that it is smooth and compatible with the topology of the space. These choices lead to different geometries in the sense of distances, geodesics etc. In the case of the S^2 example we can make the 'natural' choice of metric as we did just before by inducing it from the standard everyday metric in the space R^3 in which we have embedded the two-sphere. Squashing the sphere would naturally change the metric. What makes that choice natural is that our visual intuitions on vectors and path-length and angles still make complete sense.

Shortest distances: geodesics. We now are in a position to answer the question of what the shortest path between two points on a sphere is. That will again be the path along which photons and free particles living on S^2 would travel. Just like the route your child would presumably take on their way to the nearest two-dimensional ice-cream parlour. We will answer this question in more detail later, but let us first get a feeling and an intuitive idea of the solution. In Figure I.2.31 I have marked two points A and B on the surface, and drawn various paths between them, each of them corresponds to a segment of a circle on the surface. It is evident from the drawing that the bright yellow connection in the middle is the shortest, and it corresponds to a segment of the equator. The other paths are also segments of circles, but what sets the yellow one apart is that it is a segment of a 'great circle,' a circle of maximum size on the sphere whose radius equals the radius of the sphere itself. Great circles are defined as the intersection of a plane through the centre of the sphere (the point C in the figure) and the spherical surface. These great circles are so-called *geodesics* on the space S^2 , and correspond to what straight lines are on the plane, they correspond

to the trajectories of free particles like photons. Shortly we will discuss the equations that geodesics have to satisfy.

Vectors on curved spaces.

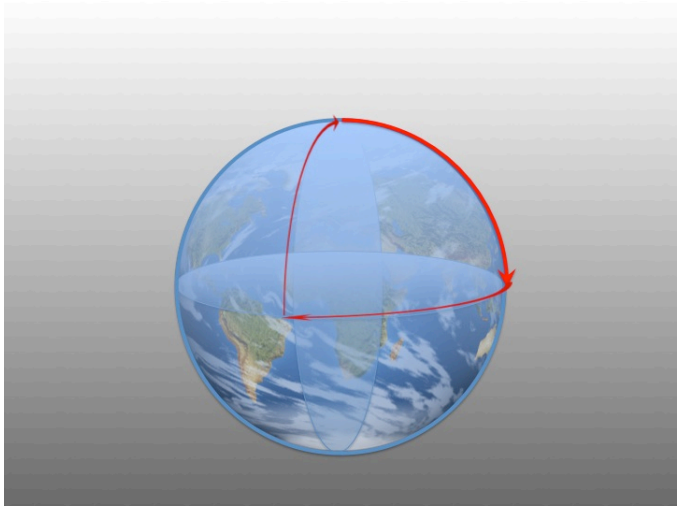
> And the curved space said: 'Vectors don't live here anymore.'

Tangent space. Assuming that physicists are also living on that sphere, they need vectors to describe what's going on: momentum vectors, forces, electric fields, and also quantum states. On a sphere, those vectors cannot live 'in the space' itself, because the sphere is not a vector space. In a curved space the notion of a 'position vector' makes no longer sense. To stay close to our flat space experience we do the following: to define a vector at some point on the sphere, we first construct the tangent plane to the surface at that point, and then put the vector there. Because the tangent space is a copy of R^2 we can add, subtract and take products of vectors there. So we attach a vector space to every point on the sphere. If we have a well-defined system orthogonal curvilinear coordinates, then locally, in any point x we can construct a set of orthonormal tangent vectors along the coordinate axes, and those define a smooth (othonormal) *frame* at any point in the patch. Having a set of smooth transition functions allows us to extend such frames over the whole manifold.

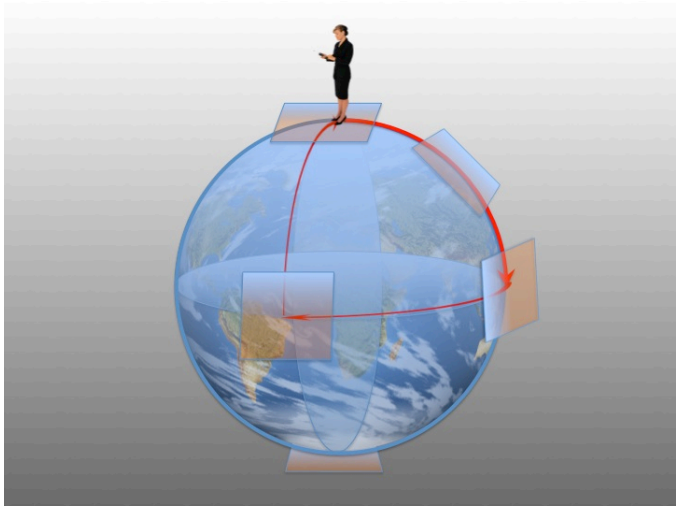
Parallel transport of vectors. Knowing how to deal with vectors at every point in space is not enough. We want to compare vectors at different points, and we want to move them around. We need to 'parallel transport' the vectors or frames from one tangent space to another. The question we are now equipped to answer is: what happens if we do the exercise with the Master Chef we did in 'flatland' before?

We put three people standing at the corners of a spherical triangle, then we draw the shortest paths between them and ask the apprentice, the messenger, to bring copies of the Chef's vector around. What happens is depicted in the Figure I.2.32 . The instruction is the same, in the point on the geodesic we first construct the vector tangent to the curve, which lies in the tangent space of the point. The Chef's vector to be transported makes a well-defined angle with the tangent vector. Parallel transport is now defined by keeping the angle between these two vectors constant while moving forward along the geodesics. The result of carefully parallel transporting a vector along the triangle is depicted in Figure I.2.32(d). On the first segment of the triangle the angle is 0 , on the second segment it is $\pi/2$, and on the third it is π . It seems to work fine, except that when the apprentice returns to the Chef, the boss is furious. It is not hard to see why. Comparing the initial and final, parallel transported vector, we see that they are not parallel at all! The transported vector has rotated over an angle of $\pi/2$. The apprentice is shocked: how could this have happened? He did after all perfectly follow the instructions all along, oh my! But the Chef is unrelenting: 'You are fired! Out through the backdoor you fool!'

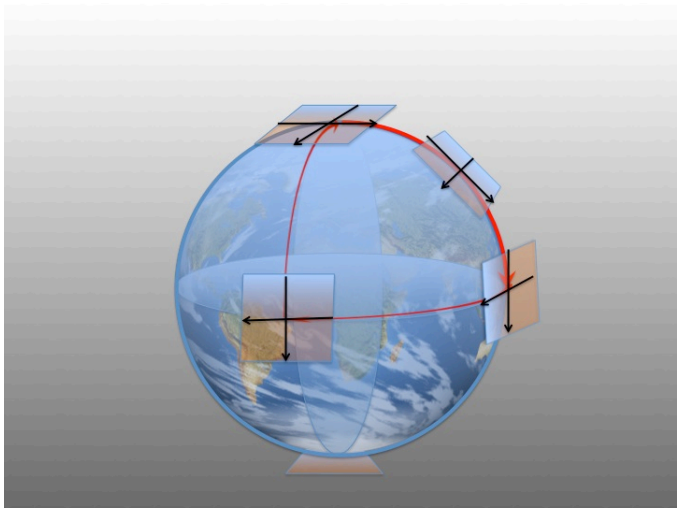
Holonomy. What we may learn from this mini-drama becomes clear when we turn the story around. It is apparently simple to find out whether you live in a curved space, without stepping outside into the embedding space; it suffices to just walk around some closed paths and parallel transport a vector along with you, and see wether it is rotated upon return. So each closed path on the manifold induces a map of the tangent space onto itself which corresponds to a rotation. This intrinsic property of a space is called *holonomy* and an important characteristic of curved spaces. For the example at hand we see that the vector is rotated by an amount that equals the solid angle bounded by the loop. The total area of a sphere is $4\pi r^2$, so 4π for the unit sphere. And indeed, the triangle covers the area (or solid angle) of an octant which equals $4\pi/8 = \pi/2$. It is easy to check that this solid angle-holonomy correspondence also holds for other simple closed paths. If we for example extend the triangle by moving the two points



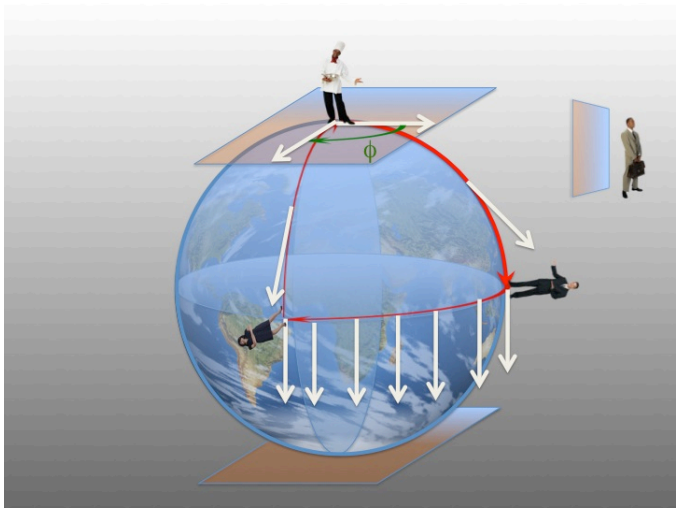
(a) Rotations generate translations on the two-sphere S^2. Triangle on curved space has more than 180^o.



(b) Vectors at a point in a curved space live in the tangent space ($\sim \mathbb{R}^2$) at that point.



(c) Carrying a frame around a triangle on a sphere.



(d) Parallel transporting a vector along a closed loop on a sphere rotates the vector.

Figure I.2.32: *The geometry underlying the tangent bundle of* S^2. Using the equivalence of S^3 with a line bundle over the two-sphere. Moving a point over S^3 is the same as carrying a tangent vector over S^2.

on the equator southwards to the South Pole, we obtain a non-trivial *two-angle* (!). Going around this two-angle will yield a holonomy of π as it should. A more general way to state the result is to say that for any closed loop in any curved space the holonomy equals the net curvature on any surface bounded by the loop, after proper normalization. As the (scalar) curvature of a sphere is a constant that equals 2, the curvature enclosed in the loop is then

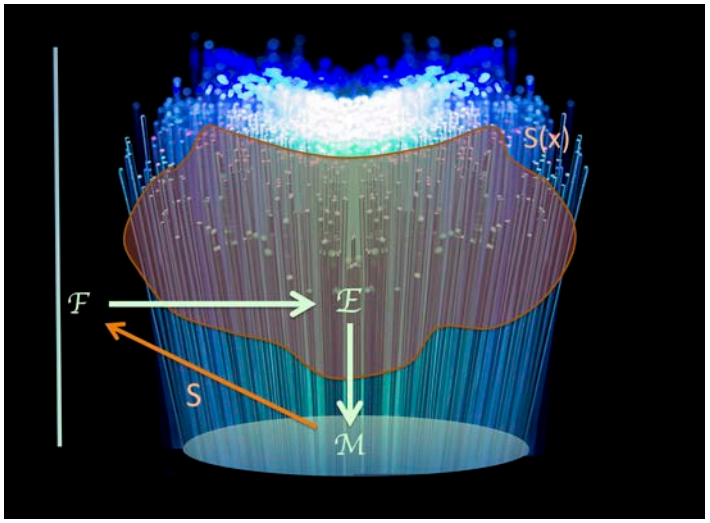


Figure I.2.33: *An optical fiber bundle.* A bundle of optical fibers, each like a finite light ray, isomorphic to the unit interval: $\mathcal{F} = [0, 1]$. The base manifold $\mathcal{M}$ is just a finite disk with a circular boundary. Every fiber can be projected down to a point in the base manifold. This bundle is 'trivial' in the sense that the bundle space is just global product, $\mathcal{E} = \mathcal{M} \times \mathcal{F}$.

equal to two times the solid angle.

Fiber bundles. In a proper description of many physical systems the most basic ingredients are some space(-time) manifold $\mathcal{M}$ that may or may not be curved, and there will be certain physical variables like temperature, a fluid flow, or whatever one is interested in. We assign functions (or fields) to those variables. For the temperature we define a function $T = T(x)$ which is a map from the base manifold to the real numbers: $T : \mathcal{M} \to \mathbb{R}$. For the velocity field this would be a vector field (also called a vector-valued function) $\mathbf{v}(x)$ which you can think of as a map $v : \mathcal{M} \to \mathbb{R}^3$.

Let us now introduce an upgraded setting for the previous paragraph, and start with a big space $\mathcal{E} = \mathcal{M} \times \mathbb{R}^n$. So the space looks very much like a bundle of fibers $F = \mathbb{R}^n$ because above any point $x \in \mathcal{M}$ we have erected a copy of the fiber. Now a function on $\mathcal{M}$ which takes its values in $\mathbb{R}^n$ can be viewed as taking a cross-section of the bundle. In other words, giving $S(x)$ corresponds to drawing some curved surface above $\mathcal{M}$ that intersects with every fiber only once. Figure I.2.33 gives an intuitive idea of such a fiber bundle. We start with a *base manifold* $\mathcal{M}$, the physical space. In this case the base manifold is a simple two-dimensional disc. Above each point of $x \in \mathcal{M}$ we erect a *fiber* $\mathcal{F}_x$ which is isomorphic to the reference fiber (drawn on the left) and in this case is a finite ray of unit length. The fibers $\mathcal{F}_x \ in \ \mathcal{E}$ are transformed images of the reference fiber. In the picture we also show local (x-dependent) map $S(x) : \mathcal{M} \to \mathcal{F}$. Such a map $S(x)$ is called a *section* of the bundle, indeed we obtain a deformed surface above $\mathcal{M}$ which is literally a cross-section of $\mathcal{E}$. In this particular example there is a smooth map from $\mathcal{E} \to \mathcal{M} \times \mathcal{F}$ from the bundle space to the global product of base and fiber, which means that the bundle is trivial.

More generally, if we think of the base manifold $\mathcal{M}$ as the space or space-time manifold, then we usually define all kinds of fields $f(x, t)$ on it. These fields often take values in some vector space $\mathcal{V}$ or an algebra, meaning that we have a map $f : \mathcal{M} \to \mathcal{F}$. A natural setting to describe both the space $\mathcal{M}$ and such a function on it is to extend the manifold to a *fiber bundle* $\mathcal{E}$, which locally for any neighbourhood $U_i \subset \mathcal{M}$ has a direct product structure $U_i \times \mathcal{F}$. The point is now that globally this is not necessarily the case. It may be that a basis cannot be extended smoothly over all of $\mathcal{M}$; In such a situation the fiber bundle framework is very powerful and versatile.

Let us illustrate this difference with another simple example. Consider the case where the base manifold M is a circle, $M = S^1$, and the fiber $\mathcal{F}$ the unit interval $\mathcal{F} = I = \{0 \leq y \leq 1\}$. The 'trivial bundle' would be a cylinder, corresponding to the global direct product $\mathcal{E} = S^1 \times I$. But we could also identify the fibers as $(\varphi = 0, y) \sim (\varphi = 2\pi, 1 - y)$, in which case we get a *Möbius band*. This band has locally the same structure as the cylinder, which means that if you only were allowed to explore your direct

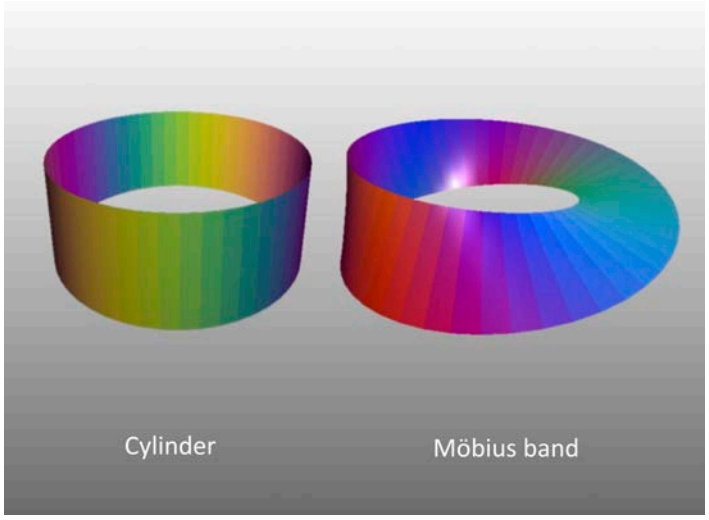


Figure I.2.34: *The Möbius band.* We have depicted the trivial bundle over the circle with the unit interval as fiber, which is a cylinder. The tangent bundle of the circle corresponds to an infinite cylinder: $\mathcal{E} = S^1 \times \mathbb{R}$. On the right we give the topologically non-trivial bundle corresponding to the Möbius band.

neighbourhood, you would not be able to decide whether you lived on a cylinder or on a Möbius band. But globally the situations are very different, walking along the inside, you end up on the outside and vice versa. In other words there is no such thing as an inside or outside as they are smoothly connected. The Möbius band is a non-orientable manifold with a single boundary. We have illustrated the trivial cylinder and the non-trivial, 'twisted' Möbius band in Figure I.2.34. The bundle picture allowed us to clearly set apart two spaces that are locally the same but globally (topologically) different. The cylinder is a two-dimensional flat space in that it needs only one coordinate patch, it has an inside and an outside separated by two one-dimensional boundaries. It is topologically like a disc with the origin taken out; it has no hole and two boundaries. If you live on the inside and your relevant-other on the outside, than that is bad news because you cannot run into each other. Remarkably, on the Möbius band that problem is non-existent.

This simple example gives a hint as to how natural and powerful the geometrical construct of a fiber bundle is. Exactly because for the base space and the fiber there are very many choices, and each of them gives rise to a subcategory of bundles with their own specific properties. You will find that there is a great variety: *line bundles, vector bundles, principle bundles, tangent bundles, frame bundles,* and many others. This world has vigorously been explored by the mathematicians, and they have developed a beautiful and rigorous framework in which many physical applications can be embedded. Books have been written about the subject and it is not our goal to delve too deeply into it, except to explore its relevance to the physics subjects treated in this book.

Tangent bundles. 

As we mentioned already, to have parallel transport and have a proper definition of distance on a curved manifold, we need extra ingredients. Having the coordinate patches with transition functions, we can draw continuous curves and parallel transport vectors. With these attributes we cannot only construct a tangent space at every point of our base manifold $\mathcal{M}$, but we can also define what is called the *tangent bundle* of $\mathcal{M}$. The tangent bundle is a smooth $2n$-dimensional manifold $\mathcal{E}$, which consists of $\mathcal{M}$ and all its tangent spaces. It has the structure of a *fiber bundle*, because above every point x of the base manifold $x \in \mathcal{M}$ of dimension n, we have erected a fiber $\mathcal{F}$ which is a copy of the tangent space $\mathbb{R}^n$. This bundle itself is a smooth manifold of dimension $2n$. For flat space $\mathcal{M} = \mathbb{R}^n$ the bundle space would just be $\mathcal{E} = \mathbb{R}^n \times \mathbb{R}^n = \mathbb{R}^{2n}$. And as we saw for the circle, the tangent bundle is just a two-dimensional (infinite) cylinder: $\mathcal{E} = S^1 \times \mathbb{R}$, it is the global direct product and therefore a trivial bundle.

The S^2 tangent bundle. The construction of the tangent bundle of S^2 is more complicated because it is topologically non-trivial. The two-sphere and various local tangent planes are sketched in Figure I.2.35, where we have also

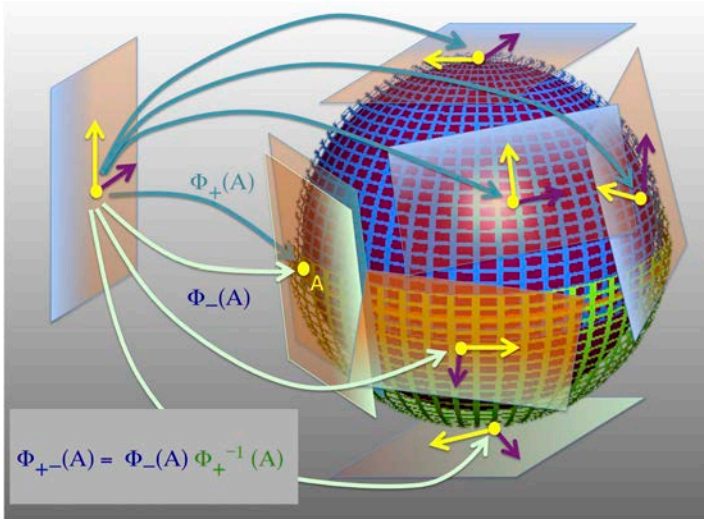


Figure I.2.35: *Tangent bundle of* S^2. The *base manifold* is the two-sphere. The *fibers* are just copies of $\mathbf{R}^2$. We have indicated the *ortho-normal frames* which are rotated with respect to a reference frame on the left. The *structure group* of the otho-normal frame bundle is just the group of rotations in the two-dimensional plane, denoted as $SO(2)$. We have indicated the transition map from the tangent frames referring to the two patches in the point $x = A$.

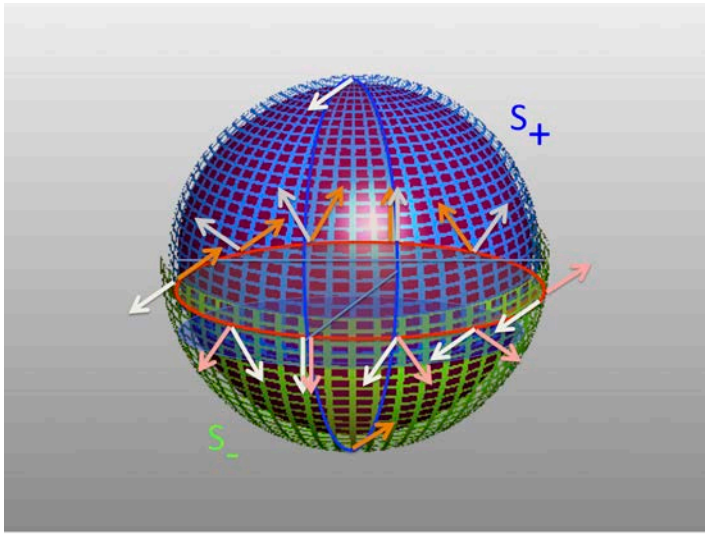


Figure I.2.36: *Transition map of coordinates and frames.* The two-sphere covered by two coordinate patches $S_\pm$, We have parallel transported a vector from the North Pole along the light blue meridians in S_+ and from the South Pole in S_- to points on the equator. Going around the equator we see that the white vectors rotate clockwise and the pink vectors anti-clockwise by an angle $\beta_\pm = \pm\varphi$. This yields the transition function $\Delta = \beta_+ - \beta_- = 2\varphi$. The topology of the tangent bundle is therefore non-trivial and has winding number $m = 2$.

shown how each fiber is related to the standard plane by some map $\Phi(\theta, \varphi)$. This map basically tells you how the basis for the fiber as vector space in each point on the sphere is rotated with respect to some reference frame. So Φ corresponds to the angle by which the frame is rotated. It means that in general in this construction of the tangent (or simpler: the related otho-normal frame bundle) there is always a rotation group involved. This map is smooth on each patch, and one obtains the transition function to go from the frame for S_+ to one for S_- at a point x in the overlap, by applying the product map $\Phi_{+-}(x) = \Phi_-(x)\,\Phi_+(x)^{-1}$. Now why is this bundle non-trivial? This is the question we turn to next.

To find out whether the bundle is trivial we focus on the transition map or gluing function in the overlap region. The result is depicted in Figure I.2.36. We start by choosing the white vector (but think of it as a frame) on the North Pole and transport it along the meridians down to the equator, there the transported white vectors are found to make an angle $\beta_+(\varphi) = \varphi$ with respect to the vector at $\varphi = 0$ (parallel transported along the equator to the tangent space at the same point). Subsequently we carry the vector at $\varphi = 0$ on its meridian all the way south, resulting in the pink colored vector at the South Pole. And from there we transport that pink vector upward along all the meridians in S_- again to the equator, yielding the pink vectors making an angle $\beta_- = -\varphi$, with the pink vector transported from $\varphi = 0$. What we have constructed is a globally smooth *section* of the frame bundle. The frames in the overlap region (the equator) on the two patches differ, and are related by a local rotation in the respective tangent

planes. We see that the transition function corresponds to a transition angle $\Delta(\varphi)$, which satisfies $\Delta(\varphi) = 2\varphi$. If we walk full-circle around on the equator once then the angle $\Delta(2\pi) = 4\pi$, has gone around twice. This means that the bundle is topologically non-trivial, because it is similar to the non-trivial twist of the Möbius band, but we here have a relative winding number $m = 2$.

Let us lift this discussion to the n-dimensional spheres S^n. The bundle is an example of a frame bundle, this bundle is linked to the group of rotations (denoted by $SO(n)$) that maps all possible ortho-normal frame choices for the tangent spaces $\mathbb{R}^n$ into each other. We cover the sphere by two overlapping disc-like patches, then the overlap is a sphere S^{n-1} times the interval. Then the transition function is a map from the overlap region into the group $SO(n-1)$. If this map is contractible, meaning that its homotopy class is trivial, then the bundle is topologically trivial.

For the two-sphere we had a transition map from the equator with coordinate φ to the frame group $SO(2)$, which is also a circle, parametrized by the angle Δ. These classes of such maps are labeled by the elements of the first homotopy group $\pi_1(SO(2)) = \pi_1(S^1) = \mathbb{Z}$, the integer $n \in \mathbb{Z}$ is often called the winding number. This means that $\Delta(2\pi) = 2\pi n$ and for the frame bundle of the two-sphere we found $n = 2$. This winding number is a topological invariant that characterizes the bundle in question. We can now also answer the corresponding question for the three-sphere, this boils down to a mapping of the two-sphere (the 'equator') into the group $SO(3)$ of three-frames. The homotopy group in question, $\pi_2(SO(3)) = 0$. So the group has only one element, which means that all the maps are contractible from which we conclude that this frame bundle is trivial. And this in turn means that the three-sphere is 'parallelizable.'

There is one other observation we want to make, which links this frame bundle of the two-sphere to the bundle that we studied in connection with the Dirac magnetic monopole. Let us recall that for the monopole we basically dealt with two vector potentials $A_\pm$ defined on two patches on the two-sphere, linked by a gauge transformation.[5] In other words we considered a map from the equator (S^1) into the gauge group of electrodynamics (which is the group phase group $U(1)$ which also corresponds to a circle: $U(1) \simeq SO(2) \simeq S^1$). We gave the explicit formula for that map $\Delta(\varphi) = \varphi$ in equation (I.1.57), meaning that the winding number for the monopole bundle equals $n = 1$.

The bundle space $\mathcal{E}$ in the monopole case corresponds to the manifold S^3, interpreted as a S^1 or phase bundle over S^2. The bundle is exactly the one described by Hopf in 1931. As we will point out in Chapter II.1, also the quantum state space of a single *qubit* corresponds to such a three sphere.

What we have learned is that the monopole and frame bundle are both realizations of a circle bundle over the two-sphere, but they are topologically distinct because they have winding numbers equal one and two respectively. The bundles with higher winding numbers correspond for example to multiply charged Dirac monopoles satisfying $eg = 2\pi n$. But the most gratifying is perhaps that in spite of their quite different physical origins these two situations could be related within the framework of fiber bundles.

Differential geometry. 

In this section we have demonstrated that for most physics applications which involve geometry we need extra structure on the manifold which allows us to properly define functions and their derivatives or integrals, and of vectors and vector fields. The structural ingredients we need are a *metric*, a *connection* or *covariant derivative*, and a definition of the *curvature* tensor. This takes us to the basic definitions of Riemannian or more generally of *differential geometry.*

[5] We talk about the concentric spherical shells for a fixed radius larger than zero.

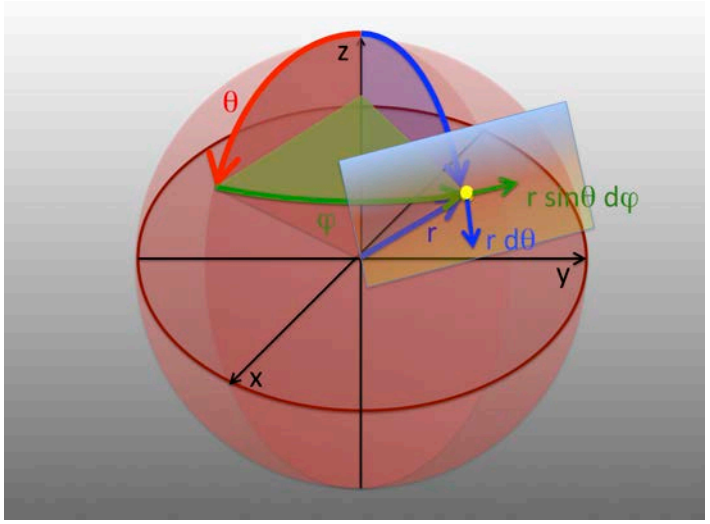


Figure I.2.37: *The geometry of the sphere.* The yellow point has spherical coordinates (r, θ, φ). The length of the equator (red circle in xy-plane) on the sphere equals $2\pi r$. The red and blue arcs therefore have equal lengths $s_r = s_b = r\theta$; with angle expressed in radians $(2\pi\ radians = 360^o)$. The green-colored segment of a horizontal spherical disc with radius $a = r\sin\theta$. For the length of the green arc follows $s_g = \varphi r \sin\theta$.

Metric. We introduced the metric with the definition of the line element in equation (I.2.16), which in general reads:

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu \qquad (I.2.17)$$

The metric is a symmetric 'tensor' with two indices which you can think of as a matrix $g_{\mu\nu}$ depending on $\mathbf{x}$,

The metric also gives a local definition of the length $|v|$ of a vector v^μ in the tangent space at a point $\mathbf{x}$ as follows:

$$|v|^2 = \mathbf{v}\cdot\mathbf{v} = g_{\mu\nu}v^\mu v^\nu = v_\mu v^\mu\,, \qquad (I.2.18)$$

where in the expressions we have adopted the standard convenient 'Einstein convention,' which says that if in any expression with repeated upper- and lower indices, these are automatically summed over, so, $v_\mu v^\mu \equiv \sum_\mu\ v_\mu v^\mu$.

For example on a two-sphere with radius r with coordinates (θ, φ) the standard metric has two non-vanishing components: $g_{\theta\theta} = r^2$ and $g_{\varphi\varphi} = r^2\sin^2\theta$. The line element ds follows from:

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu = r^2(d\theta^2 + \sin^2\theta d\varphi^2) \qquad (I.2.19)$$

Looking at Figure I.2.37 it is not hard to see why. You may verify that, (i) for φ fixed the arc or path length on the sphere corresponding to an angular displacement $d\theta$, corresponds to $r\,d\theta$ (as θ runs along a big circle), (ii) for fixed θ, the path length corresponding to the angular displacement $d\varphi$, equals $r\sin\theta\,d\varphi$ (as the φ variable runs along a 'lateral' circle with radius $r\sin\theta$). What this means is the following: if we change the coordinate φ for an arbitrary point on the sphere by an infinitesimal amount $d\varphi$ then the length of the corresponding displacement vector is $ds = r\sin\theta d\varphi$. So the metric tensor locally links infinitesimal changes in coordinates to infinitesimal path lengths in the space.

Path length. The length L_{ab} of the curve $\gamma(t)$ parametrized by t, between two points $\gamma(a)$ and $\gamma(b)$ is naturally defined as the integral:

$$L_{ab} \equiv \int_a^b |v(t)|dt\,, \qquad (I.2.20)$$

in a different more familiar wording, the distance traveled is just the magnitude of the velocity component along the trajectory integrated over the appropriate time interval. So for example if we choose the lateral green circle (with θ constant) in Figure I.2.37 , we would have $\gamma(t) = \{\theta, \varphi(t)\}$:

$$L = \int r\sin\theta\frac{d\varphi}{dt}dt = r(\varphi(b) - \varphi(a))\sin\theta\,, \qquad (I.2.21)$$

as it should.

Frames. We like to mention that there is a slightly different formulation for dealing with Riemannian geometry due to Cartan. This formulation is close to the standard form in which gauge theories of the non-gravitational interactions are presented. We start by introducing an ortho-normal

basis or frame in the tangent space by writing:

$$g_{\mu\nu} \equiv \eta_{ab} e^a_\mu e^b_\nu \,, \tag{I.2.22}$$

where η_{ab} is the usual flat space metric (or inner-product), and the funny objects e^a_μ are the so-called *solder forms* or *'vielbeine'* that convert vectors from the curvilinear coordinate components to the 'flat' components. So for the spherical surface we could simply choose $e^1_\theta = r$ and $e^2_\varphi = r\sin\theta$ and all others equal zero. These define what is called a local orthonormal frame $\{e^a\} \equiv \{e^a_\mu \, dx^\mu\}$.

With these definitions the inner product can be rewritten as:

$$\mathbf{v}\cdot\mathbf{w} = g_{\mu\nu} v^\mu w^\nu = \eta_{ab} e^a_\mu e^b_\nu v^\mu w^\nu = \eta_{ab} v^a w^b \,, \tag{I.2.23}$$

with the flat space vector components defined as $v^a \equiv e^a_\mu v^\mu$.

Connection. Given the metric g or the frame $\{e\}$ we define the so-called metric connection ω, which written in components would read $\omega_\mu{}^a{}_b$, meaning that it is like a space(time) (row) vector and acts like a matrix in '$a-b$' space. This connection is defined by the linear set of equations:[6]:

$$de + \omega \wedge e = 0 \,. \tag{I.2.24}$$

Specifying the metric, the metric connection or the set of 'vielbeine' are equivalent characterizations of the manifold. Knowing the frame $\{e\}$, one can solve equation (I.2.24) for the connection in terms of the vielbeine and their derivatives. For the two-sphere the result for the connection is simply $\omega_\varphi{}^1{}_2 = -cos\theta$. Note that it has two flat indices and therefore it acts like a matrix in flat space. We introduce the connection ω_μ, because it is similar to the gauge potential A_μ in gauge theory. The gauge transformations in the case of general relativity would correspond to *local* orthogonal (or Lorentz transfomations) rotations of the frame that leave the metric in other words the angles and lengths of vectors invariant,

$$e'^a = \Omega^a{}_b e^b \,. \tag{I.2.25}$$

Curvature. To complete this lightning review of non-Euclidean or Riemannian geometry, we have to add a final ingredient, which is the Riemann *curvature tensor* or two-form R, which is the strict analogue of the 'field strength' F in gauge theories. It can be calculated from the connection as follows:

$$R = d\omega + \omega \wedge \omega \,. \tag{I.2.26}$$

This Riemann curvature is an object with four indices. We will refrain from descending any further in this myriad of indices except for at least giving the result for the two-sphere. There is basically only one component that is non-zero: $R^1{}_2 = R^1{}_{212}\, e^1 \wedge e^2 = \frac{1}{r^2}\, e^1 \wedge e^2$. From this Riemann curvature one finds the Gaussian curvature as $R_G \equiv R_{abab} = 2/r^2$. We say that the Gaussian curvature of the sphere is constant. It does not depend where you are on the sphere, it only depends on the radius of the sphere. If that radius becomes large the curvature tends to zero. In other words the space becomes effectively flat.

The main point of this subsection is to show that the analytical structure of differential geometry is highly canonical. It involves three subsequent defining equations: it involves three subsequent defining equations: (i) for the metric (I.2.22) or the frame, (ii) for the connection in terms of the frame (I.2.24) and (iii) for the curvature in terms of the connection (I.2.26). Our aim is *not* to make any real computations but merely to get across that at this level of analysis it is evident that general relativity and gauge theories share an underlying geometric structure. Roughly stated, both involve a connection and a curvature defined in terms involving derivatives of the connection.

The geodesic equation. Geodesics are the paths along which free particles move. We have asked what the shortest path between two points is on a sphere and found it

[6] We use the quite compact index free notation because it makes the underlying structure more transparent. With indices the above equation would look quite daunting: $\partial_\mu e^a_\nu - \partial_\nu e^a_\mu + \omega_\mu{}^a{}_b\, e^b_\nu - \omega_\nu{}^a{}_b\, e^b_\mu = 0$.

to be a segment of a great circle. In general that question can be answered by minimizing the path length L_{ab} under variations of the path. From the metric one can directly construct a free particle Lagrangian, which is a function of the coordinates and their time derivatives:

$$\begin{aligned}\mathcal{L}(x^\mu, \frac{dx^\mu}{dt}) &= \frac{1}{2} m g_{\mu\nu} \frac{dx^\mu}{dt}\frac{dx^\nu}{dt} \\ &= \frac{1}{2} m r^2 \left((\frac{d\theta}{dt})^2 + \sin^2\theta (\frac{d\varphi}{dt})^2 \right) \\ &= \frac{1}{2} m r^2 (v_\theta^2 + v_\varphi^2) \end{aligned} \qquad \text{(I.2.27)}$$

Minimizing the time integral of $\mathcal{L} \sim |v|^2$ (instead of L_{ab}) one obtains the so-called Euler-Lagrange equations. On the two-sphere one obtains:

$$\begin{aligned} &\frac{d}{dt}(\frac{d\theta}{dt}) - \cos\theta \sin\theta (\frac{d\varphi}{dt})^2 = 0\,, \\ &\frac{d}{dt}(\sin^2\theta \frac{d\varphi}{dt}) = 0\,. \end{aligned} \qquad \text{(I.2.28)}$$

These are the Newtonian equations of motion for a particle on a sphere in the absence of an external force as discussed in the section on Newtonian mechanics in Chapter I.1. All terms have two time derivatives, among them are the pure 'accelerations' in the θ and φ directions. There is no potential as such and the extra terms that appear are a consequence of spherical geometry. So like in flat space, where a force would typically curve the orbit, and straight lines (describing shortest distances) are obtained by setting the force equal zero, something similar is true in curved spaces where free particles move along *geodesics* and they do independently of their mass or momentum.

Let us check a few simple solutions. For example, if we assume that the velocity component in the φ direction, $\sin\theta d\varphi/dt$ vanishes, we obtain the solutions $d\theta/dt = constant$. These describe a particle moving with constant velocity along any meridian (where $\varphi = constant$). This shows that the meridians are indeed shortest paths. Choosing $d\theta/dt = 0$, on the other hand, gives the solution, $\theta = \pi/2$, $d\varphi/dt = constant$, which corresponds to the particle moving with constant velocity along the equator, again a 'big' circle or geodesic.

These calculations confirm our previous observations with respect to the Figures I.2.31 and I.2.32, where we saw that the shortest path between two points is always a segment of a great circle. That allowed us to also draw a triangle on the sphere as we did in Figure I.2.32, and what we see is that the triangle has three 90^o angles. In other words that the sum of the three angles of this triangle is 270^o which is far more than the 180^o of a planar triangle. It is amusing to note that if you move the two lower points of the triangle to the South Pole, you get a non-trivial 'two-angle.'

Let us finally note also that all shortest paths from the North to the South Pole, in other words all meridians, are in fact 'parallel,' because they all are perpendicular to the equator. Indeed, in a curved space 'parallel lines' may cross. Boy! Yet another reason why life on Earth is so complicated. My devise would be, be prepared: think global and act local, rather than the other way around. ■ ■

The geometry of gauge invariance

A gauge theory is the prototype model for all fundamental interactions, where the gauge field may either describe the *electromagnetic field* corresponding to the *photon*, or the fields mediating the strong interactions corresponding to 8 *gluons*, or the weak interactions described by the $W^\pm$ and Z *bosons*, and finally it may describe general relativity corresponding to the gravitational interaction mediated by the *graviton*. The gauge symmetry principle is therefore a fundamental and universal hallmark of nature. Gauge invariance imposes a strong constraint on the system of fields involved. In particular it completely fixes how the *force carrying fields* just mentioned interact with the *'charge' carrying fields* or constituent particles like the the electon, the muon, the neutrinos or the quarks. On the other hand this

physically based gauge principle is deeply linked to the geometry of fiber bundles of which the tangent bundle we discussed is only one example.

The topic of gauge invariance will pop up in this book at regular intervals. Here we give some of the mathematical background of the gauge principle which corresponds to the geometry of fiber bundles. In Chapter I.4 on the quest for the basic building blocks of matter we discuss how the gauge theory approach has led to the standard model of the fundamental interactions between elementary particles. And in Chapter II.6 on symmetries and their breaking we describe in more detail what the equations for gauge theories look like and how they implement the idea of a local gauge invariant dynamics.

The charge degree of freedom. To get a better feeling for the notion of charge and its connection to gauge invariance it may help to explicitly introduce a model for electric charge. Think of particles carrying an extra periodic coordinate β that labels points in some 'internal space' that in this case corresponds to a tiny circle. You may think of β in that sense as an extra charge degree of freedom, and the charge $q = ne$ as a kind of momentum in this internal electromagnetic dimension with coordinate β. The particle carries along a charge-phase factor

$$f_n(\beta, x) = e^{in\beta(x)} . \qquad \text{(I.2.29)}$$

If we split this phase factor in its real and imaginary parts by writing it as $\cos n\beta + i \sin n\beta$, we represent the phase factor as a little two-dimensional unit vector making an angle $n\beta$ with the real (horizontal) axis.[7] Note that if we vary β from 0 to 2π, then the phase of the particle with charge number n changes by $2\pi n$, so the corresponding little vector rotates n times as fast.

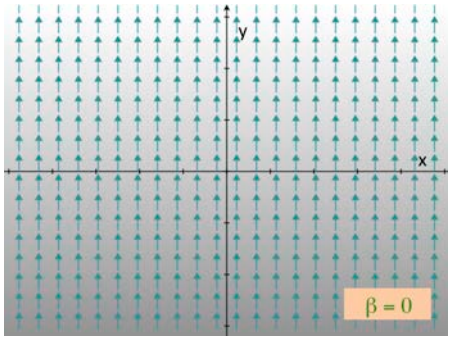
You may say that the charge-number corresponds to the 'momentum" in the β direction because it is proportional to the beta derivative $-ie\partial_\beta f_n = q\, f$, and as there is no β-dependent potential or force, the β-momentum (= charge) is conserved. The dynamics in beta-space is therefore entirely trivial and that is precisely why nobody talks about it in the first place. But it at least explains the terminology that charge corresponds to an *internal degree of freedom*. I think that it is also quite helpful for getting a better understanding of the notion of gauge invariance. And moreover, if we would treat this little charge-degree of freedom as a quantum particle on a circle, the momentum (= charge) would be quantized as well. It would look like the Bohr model applied to the quantization of charge.

Gauge transformations. The best way to think about (local) gauge transformation is as a position- and time- dependent rotation, not in real space but in some internal vector space, that is carried by each of the matter fields. To clarify this let me return to electromagnetism. Another way to look at the local charge-phase factor we introduced is that it is the phase of a field $\Phi(x) = f_n(x)\phi(x)$ having a charge $q = ne$. In quantum theory a particle with charge $q = ne$ is described by a complex wavefunction $\Phi(x) = f_n(x)\phi(x)$, and $f_n(x)$ represents the local phase of that wavefunction and the function $\phi(x)$ its magnitude. Formally a gauge transformation acts on the fields (in fact on the phase factor) as follows:

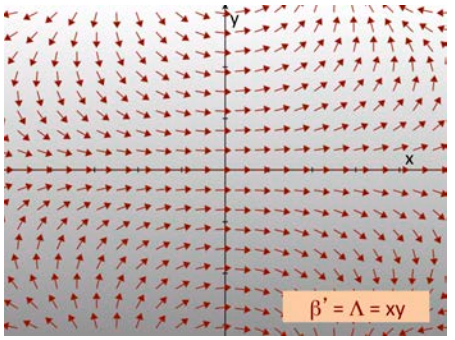
$$f_n \to f'_n = U_n f_n \ \text{ with: } \ U_n(\Lambda) \equiv e^{in\Lambda(x)} . \qquad \text{(I.2.30)}$$

The transformation U_n corresponds to a unitary representation of the group $U(1)$ labeled by the integer $n \in \mathbb{Z}$. It is unitary because $U^*U = U^{-1}U = \mathbf{1}$. And the gauge group of the theory is therefore naturally called $U(1)$, because a phase factor can be thought of as a (1×1) unitary matrix. If you prefer to talk about the little vector, then you should refer to the gauge group $SO(2)$, the group of rotations in the two-dimensional plane, but that group is the same as (or isomorphic to) $U(1)$ because its elements are also labeled by an angle.

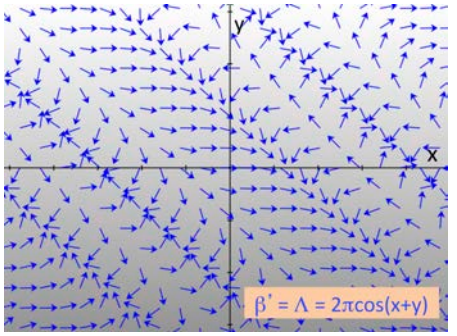
[7] If you are unfamiliar with the notion of a complex phase factor you might want to look at the *Math Excursion* on complex numbers at the end of Part III on page 630.



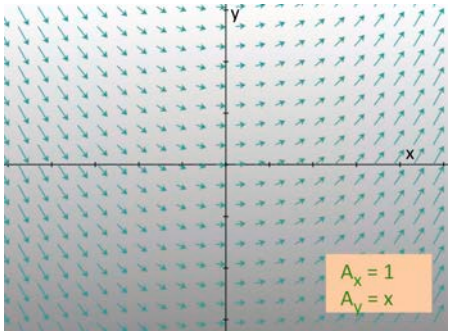
(a) Charge-phase factor f_1 in the trivial gauge $\beta(x,y) = 0$.



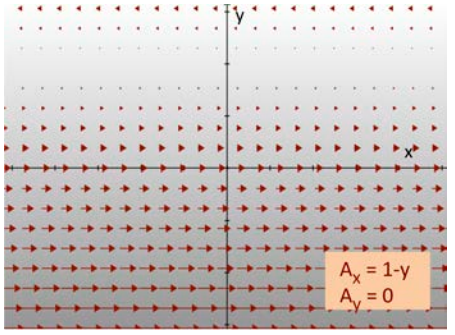
(b) Charge-phase factor f_1 in the gauge given by $\Lambda = xy$.



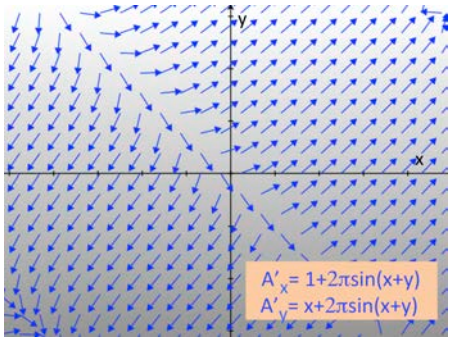
(c) Charge-phase factor f'_1 in the gauge given by $\Lambda = 2\pi\cos(x+y)$.



(d) The gauge potential yields $\mathbf{A} = (1, x)$, yields a constant magnetic field in the z-direction.



(e) The vector potential after the transformation $U = e^{i\Lambda}$. With $\mathbf{A}' = \mathbf{A} - \boldsymbol{\nabla}\Lambda$ with the Λ given above.



(f) The vector potential in a gauge with Λ given above.

Figure I.2.38: *Gauge transformations* The effect of two different local gauge transformations on the phase factor $e^{i\beta(x,y)}$ and on the gauge potential $\mathbf{A} = (A_x, A_y)$. All describe the same uniform magnetic field that is directed out of the page to the front.

So, properly speaking, the phase factor $f_n(x)$ of the wavefunction is an element of a one-dimensional complex, or two-dimensional real vector space $\mathcal{R}ep$, on which the unitary representation of the gauge group $U(1)$ with label n acts as a transformation. In brief, if I make a gauge transformation $\Lambda(x)$, the phase factor of the field will transform by multiplication with a phase factor $\exp\left(in\Lambda(x)\right)$ and therefore its phase gets shifted by $n\Lambda(x)$. And the gauge potential transforms like indicated in the formula (I.1.48).

What gauge invariance means is that we are free to choose a frame of basis for the two-dimensional vector space in which unit charge vector $f_n(\beta, x)$ lives, at every point x independently. Very much like the tangent spaces we discussed before. In other words at any point x we have the choice of which point on the circle we call the origin to which we assign the value $\beta = 0$. *Gauge invariance is the statement that the physics does not depend on that local choice.* This implies that the physics doesn't change if we shift β at each point x by an amount $\Lambda(x)$. We have illustrated this in Figures I.2.38, where we have depicted both the phase $\beta(\mathbf{x})$ and $\mathbf{A}(\mathbf{x})$ in three different gauges for a situation in two spatial dimensions. These images

underscore the great generality of local gauge transformations, and in spite of looking so different, the magnetic field $\mathbf{B} = \boldsymbol{\nabla} \times \mathbf{A}$ is the same for all three configurations. It is a gauge invariant quantity and corresponds in this case to a uniform field in the positive z-direction.

The reasoning above clarifies the use of the term 'gauging.' If you think of a little pointer moving over a dial, then applying a gauge transformation amounts to redefining the label 'zero' on the dials located at different positions x. It is like calibrating the dials in all of space-time.

Gauge covariant derivative. Let us now consider the following question. I have a space with some electromagnetic fields (or potentials) in it, and I take a charge that sits at $x = x_0$ and move it slowly along some path γ to another point x_1. What will happen? I was careful enough to not disturb the fields, but as the charge interacts with the fields along the way, did something happen to that phase perhaps? Well, certainly, and what will happen is that the phase β will change on its way to x_1. By what amount does it change? And does that change depend on the path I choose? These are the questions that we turn to next.

Let us take small steps at a time, or better even, infinitesimal steps! So, suppose we want to know what the charge-phase would look like at a nearby point, then we can make a linear approximation only keeping the first derivative:

$$f_n(x') = f_n(x + \triangle x) \simeq f_n(x) + \triangle x \frac{df_n}{dx}\Big|_x + \cdots , \quad \text{(I.2.31)}$$

but this does not take care of the change in the frame in which the phase is expressed, by which I basically mean the orientation of the real and imaginary axes of f_n at different points x. That basis change is determined by the gauge connection $A_\mu(x)$, which means that we have to replace the ordinary derivative with the so-called *gauge covariant derivative*:

$$\partial_\mu \to D_\mu \equiv \partial_\mu + iqA_\mu . \quad \text{(I.2.32)}$$

The added gauge connection literally connects the frames in neighbouring points. It is not sufficient to just calculate the phase; in order to compare the phases at different points you need to know how the bases at those points are related.

Why the term covariant derivative? It is like the derivative in a co-moving frame, and therefore the appropriate term indeed. This becomes clear if we look at how this derivative transforms under gauge transformations given that the field transforms as given in (I.2.30) and the potential like that given in (I.1.48) as $A_\mu \to A'_\mu = A_\mu - \partial_\mu \Lambda$. We find:

$$D_\mu f_n \to [D_\mu f_n]' = (\partial_\mu + iqA'_\mu) f'_n = U_n [D_\mu f_n] . \quad \text{(I.2.33)}$$

which shows that this derivative transforms covariantly indeed, in other words, just like the f_n itself. This is an important observation because one sees that quantities like $|D_t f_n|^2$ and $|\mathbf{D} f_n|^2$ will be gauge invariant and these are terms that appear in the expression for the energy density of the field f_n. And this in turn implies that to get an invariant energy the interaction between the charged field (or particle) has to be of a form involving the gauge-covariant derivative. That is what it means to say that gauge symmetry dictates the detailed nature of the interactions!

Suppose we have a function $h(x)$, imposing that $dh/dx = 0$ implies that $h = constant$. Something similar can be defined for the covariant derivative. The equation for what is called a *covariantly constant* charge vector reads

$$D_\mu f_n = 0 . \quad \text{(I.2.34)}$$

The solution to this equation amounts to expressing a path dependent relation between the phase at two points, corresponding to the parallel transport of the charge-phase along a given curve. Let us look at this statement in more detail.

The gauge connection. To carry the phase factor around we need a somewhat fancy expression involving the gauge

connection A_μ. Let me recall the line integral $I(\gamma; x_0, x_1)$ of the gauge field along a curve γ given in (I.1.51) and depicted in Figure I.1.26. To parallel transport the phase along some curve γ going from x_0 to x_1, we need to use not just the line integral $I(\gamma; x_0, x_1)$, but rather its exponential:

$$W_n(\gamma; x_0, x_1) \equiv e^{i\,n\,I(\gamma; x_0, x_1)}\,. \qquad \text{(I.2.35)}$$

This path dependent phase factor carries exactly the frame from x_0 to x_1 so that:

$$f_n(\beta, x_1) = W_n(\gamma; x_0, x_1)\, f_n(\beta, x_0)\,. \qquad \text{(I.2.36)}$$

This expression furnishes the general solution to the equation (I.2.34) for the covariantly constant charge-phase f_n. It transports the phase in a gauge covariant way, that is to say in such away that under a gauge transformation we have that

$$W_n \to W'_n(\gamma; x_0, x_1) = U_n(x_1) W_n(\gamma; x_0, x_1) U_n^\dagger(x_0),$$

and this means that in the equation (I.2.36), the combined effect of a gauge transformation on all factors is the same on the left- and right-hand side. The net effect is a multiplication by $U_n(x_1)$ from the left, as it should be according to (I.2.30). So we have answered both questions: how the charge-phase will change and that it does so depending on the path chosen. The gauge 'connector' is nothing but the path dependent phase factor W_n.

Gauge theory and principal fiber bundles. The central concept describing both the gauge potentials and the underlying space-time manifold $\mathcal{M}$ is called a *principle bundle* denoted by $\mathcal{E}$, consisting of a space which locally can be thought of as a direct product of the *gauge group* G and the base manifold $\mathcal{E}$: $G \times \mathcal{M}$.

In Figure I.2.39 we have given a simple example in which the base manifold is a circle $\mathcal{M} = S^1$ parametrized by an angle φ (the red circle). For the group we have chosen the group of rotations in the plane denoted by $SO(2)$, parametrized by an angle Λ making the group space also a

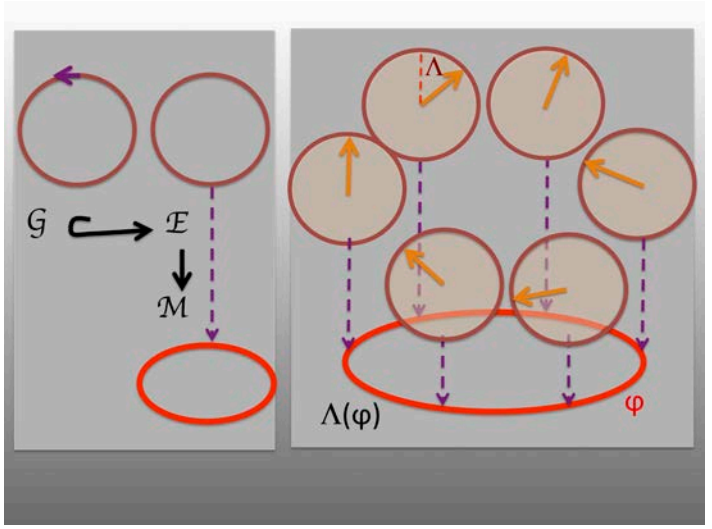


Figure I.2.39: *A principle bundle associated with a gauge group* $\mathcal{G}$. Here the group is the phase group $U(1)$, which is a circle (in brown) parametrized by $0 \leq \Lambda < 2\pi$. The base space $\mathcal{M}$ is also a circle (in red) with $0 \leq \varphi < 2\pi$. Above each point in the base space we have a fiber that is a copy of the group $\mathcal{G}$ labeled by an angle Λ. Choosing a gauge corresponds to choosing a (cross) section of the bundle: a particular choice for $\Lambda = \Lambda(\varphi)$ as indicated on the right.

circle $\mathcal{G} = S^1$ (the purple circle). This group is the same as the phase group $U(1)$, and as we saw this group is actually the gauge group of electrodynamics. Above each point of $\mathcal{M}$ we have a copy of $\mathcal{G}$ with an angle $\Lambda = \Lambda(\varphi)$.
The point is now that any electromagnetic field configuration corresponds to a particular bundle, and choosing to write down the configuration of the potentials explicitly we have to 'choose a gauge' which amounts to choosing a particular cross-section through the fibers specifying a particular choice of $\Lambda = \Lambda(\varphi)$. And this is for example done in the picture on the right-hand side.

Charge carrying fields and associated bundles. Often the gauge field is called the gauge connection, because it connects local coordinate frames at different points with each other. In general a charge carrying field carries a *representation* of the gauge group and these correspond to so-

called *associated bundles*, where we have attached a copy of the representation space Rep of the group $\mathcal{G}$, to every point of the base manifold in some smooth way.

Returning to our previous example a field carrying a charge $q = ne$ would be described by a complex field say $\psi_n(x)$, which has a magnitude and a charge-phase that again may depend on position, so,

$$\psi_n(x) = \exp(in\beta(x))\rho(x)\,.$$

Gauge transformations on such a field act as a local phase transformation; we multiply the field with the local phase factor $U_n\big(\Lambda(x)\big)$:

$$\psi_n(x) \to \psi'_n(x) = U_n\big(\Lambda(x)\big)\psi_n(x)\,. \qquad \text{(I.2.37)}$$

Examples are depicted in the Figures I.2.39, and I.2.40, which give you an impression of the case where $\mathcal{M} \simeq S^1$, $\mathcal{G} \simeq U(1)$. The representations act on the $f_n(\beta(\varphi))$ and the representation space can be depicted by the little charge vector. In Figure I.2.39 you see that the fiber corresponds to the orbit of the charge vector under rotations, and a specific bundle is obtained by choosing a particular gauge which means that above every point of $\mathcal{M}$, you choose a particular vector making sure that the overall configuration is smooth. This is appropriately called *choosing a (cross) section* of the bundle. This leads to for example the configurations depicted in Figure I.2.40 of a number of smooth closed ribbons. The configuration on the left represents the constant phase $\beta(\varphi) = 0$, corresponding to the connection $A = 0$ of the trivial bundle. The other phase configurations are smooth deformations that correspond to gauge transformations. So all three represent the same physical situation in different gauges. They are *gauge equivalent* configurations. It clearly demonstrates the local character of the gauge transformations, because at any point of the base manifold we can choose a different rotation, as long as the overall deformations correspond to smoothly 'wiggling' the configuration.

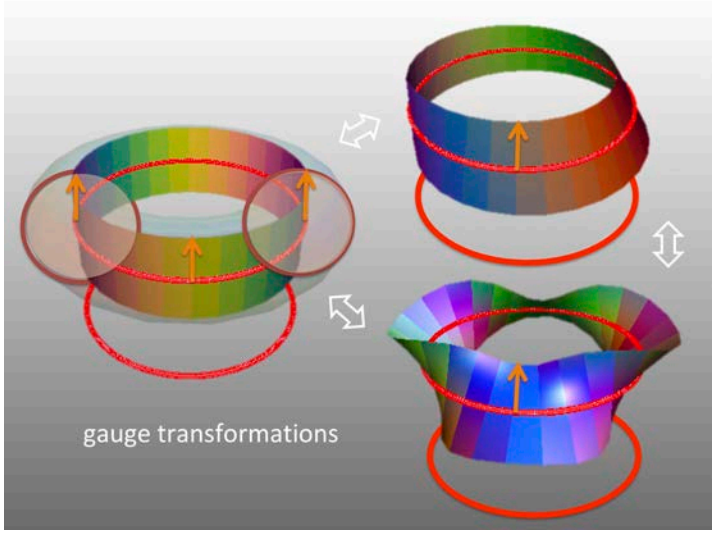


Figure I.2.40: *Gauge equivalence.* Here we have depicted gauge equivalent configurations of the charge-phase factor $e^{in\beta(\varphi)}$ on a circle. On the left we have the *trivial* configuration $\beta(\varphi) = 0$. Gauge transformations correspond to 'wiggling' the configuration. The configurations above are related by a periodic transformation $\Lambda(\varphi)$ so that $\beta \to \beta' = \beta + \Lambda$, with $\Lambda(0) = \Lambda(2\pi)$.

Gauge invariant characteristics. You may wonder if it is possible in reality to 'drag' a state vector along a closed loop like we described and whether the resulting phase change can be measured. The answer is yes and the fact is that what I have described is known as the *Berry phase* after the British mathematical physicist who discovered that it was possible to identify it in certain setups with time- or space-dependent Hamiltonians. The effect is also closely related to the much older *Aharonov-Bohm effect* as will be discussed in Chapter II.3. This interference pattern depends on the solid angle that the path $H(\lambda)$ has covered on the sphere.[8] Interestingly the Berry phase is apparently a purely geometric phase depending only on the geometry of the space of Hamiltonians.

[8] The path is oriented and the orientation decides whether to take the solid angle ω or $4\pi - \omega$, which with equation (II.3.4) amounts to $R_k(\theta) \to R_k(-\theta)$.

Other gauge groups. We have in this subsection exhibited the structure of a principal fiber bundle and the vector bundles associated with its representations, but only for the rather modest example of the group $\mathrm{U}(1)$. This may have come across as a demonstration of how to crack a peanut with a sledgehammer. We want to stress that the fiber bundle description is quite universal as it is the framework in which most classical physics involving local symmetries can be cast. For example the *Standard Model* involves $\mathrm{U}(1)$, $\mathrm{SU}(2)$ and $\mathrm{SU}(3)$ gauge fields describing the electromagnetic, weak and strong interactions respectively. And the *Grand Unified Theories (GUTs)* that we will discuss in Chapter I.4 have even larger gauge groups like $\mathrm{SU}(5)$ or $\mathrm{SO}(10)$ involving more gauge interactions.

It means that the fields take a value in a *representation space* which is a vector space $\mathcal{V} = \mathcal{R}ep$ which is typically $\mathbb{C}^n$ or $\mathbb{R}^n$. And the corresponding unitary representation of the group works in this space as a linear transformation (say, a rotation). The group can be any compact group like the groups of unitary or orthogonal $(N \times N)$ matrices denoted by $\mathrm{SU}(N)$ or $\mathrm{SO}(N)$ respectively. The label n on the field refers to the dimension of the vector space on which some irreducible (unitary) matrix representation of that gauge group acts. The *Math Excursion* on groups on page 635 of Part III gives a basic introduction to group theory. As we saw the group $\mathrm{U}(1)$ is special in that all representations are one-dimensional, meaning just phase factors.

For the group $\mathrm{SO}(3)$ the unitary representations are labeled with a semi-positive integer l, where the group is then represented by $(2l+1) \times (2l+1)$ matrices. This representation acts as a transformation group on a $(2l+1)$-dimensional vector space. A field in this $\mathrm{SO}(3)$) gauge theory will take a value in one of these vector spaces and is said to carry integer spin l. When the spin equals one we have the standard three-dimensional vector but one that lives in a $\mathcal{R}ep = \mathbb{R}^3$ internal space.

For $\mathrm{SU}(3)$, the gauge group related to the strong interactions, the quarks and antiquarks transform as 3-dimensional representations (color *triplets* and *anti-triplets*), while the gluons form an 8-dimensional representation. This indeed means that $\mathrm{SU}(3)$, the group of 3×3 unitary matrices with a unit determinant, also has a representation by 8×8 unitary matrices, which is *irreducible*, meaning that it cannot be reduced to a lower dimensional (for example three-dimensional) representation. This representation acts on the eight-dimensional vector field, which describes the gluons. A major achievement in mathematics has been that in the early twentieth century all these continuous groups and their representations were classified. The results have found a rich variety of applications in physics as we will show in Chapter I.4 where we discuss the phenomenology of the 'Standard Model.'

In the previous subsection we argued that to describe vector fields on curved spaces one needs to introduce the so-called 'tangent bundle' of the manifold. This means that also general relativity can be cast as a gauge theory where the local gauge group is the symmetry group of the local structure of space-time. Locally our space-time is flat Minkowski space-time with its translation and the Lorentz symmetries. The corresponding group is called the *inhomogeneous Lorentz* or *Poincaré group*. The field strength in that case corresponds to the local curvature tensor R of the manifold and the connection would be the so-called metric connection ω_μ that we introduced in equations I.2.26 and I.2.24. It is gratifying to see that these phenomenologically so different fundamental interactions that we have encountered in nature share this underlying structure of gauge invariance, mathematically represented by the concept of a fiber bundle. We must add the important fact though, that the physics itself resides in the field equations, being the Maxwell (more generally, the Yang-Mills) and Einstein equations. We return to the Yang–Mills equations in Chapter II.6 on symmetries and their breaking. The bundle picture makes the mathematical setting transparent and clarifies some of the physical features. ■ ■

The physics of information

> It would appear that we have reached the limits of what is possible to achieve with computer technology, although one should be careful with such statements, as they tend to sound pretty silly in 5 years.
>
> *John von Neumann (1951)*

Computation necessarily involves information storage and the manipulation of information on some underlying physical substrate, so far mostly based on semiconductor technology. Information is stored in the states of the system and one can manipulate the states by interacting physically with that system. If the scaling down of basic components is to continue as is predicted by Moore's law, then entering the quantum domain is inevitable. So, there is a quantessence to information as well. This has profound consequences for how we should think about information and information processing. It turns out that quantum computation offers fundamentally different options for tackling certain classes of hard problems.

A bit of information. Volumes are typically measured in liters, gallons, pints or cubic meters; and the unit chosen strongly depends on the local context. For information, however, this does not hold; it is universally measured in *bits*. This canonical character derives from the fact that the introduction of computers was right from the start a global affair. The 'bit' is the smallest unit of information and forms the basis for digital memories and data processing devices. One bit can be represented in many ways, for example like a switch that is on or off, or a single digit binary number being either one or zero, or equivalently as a magnetic spin pointing either up or down, or a number that is either plus or minus one (see Figure I.2.41). If I want to qualify for a discount on a public transportation ticket for example, only one bit of information concerning my age will do. I only have to answer one yes-or-no question: are you younger or older than 65? In answering a single yes-or-no question you provide one bit of information. Generally quantitative thinking is based on working with variables that can be assigned numerical values; we attach numbers to them even though these may be only approximate. Those finite approximations can always be converted to finite base-2 or *binary* numbers, only containing one's and zero's, and any calculations that you would like to do with the original numbers can also be performed in base-2. And we all know that such calculations can be extremely well and swiftly performed by today's digital devices, at least if an efficient algorithm is available.

Figure I.2.41: *The bit.* Various representations of a bit of information. It is a two-state system such as a switch, a particle that can be in either of two states, or a classical spin that can point up or down.

Information and entropy

State counting, entropy and information. In all information devices the information is carried by a physical substrate representing a certain number of bits. The amount of information that can be stored in a physical system is de-

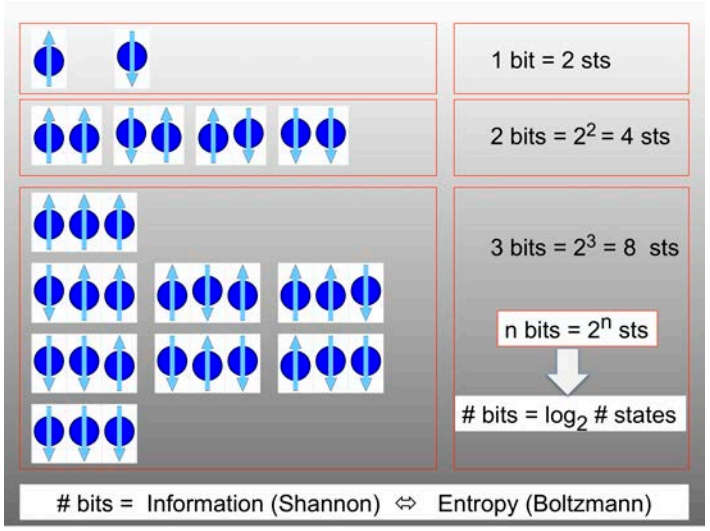


Figure I.2.42: *Entropy and information.* Counting the number of states of a digital memory. Shannon defined the *information capacity* of a system in bits as the logarithm of the number of states, therefore information is directly proportional to the notion of entropy in physics as defined by Boltzmann in the nineteenth century.

termined by the number of distinct states that the system can be in. Let us think of a memory consisting of an array of little magnets that can point either up or down. Then we can count the number of states of such an array of bits as we did in Figure I.2.42. For one bit we have 2 states, for two bits it is $2 \times 2 = 4$ states, and for n bits it is clearly $2 \times 2 \times ... \times 2 = 2^n$ states. This shows that there is a direct relationship between information capacity, i.e. the number of bits, and the number of states. This is an *exponential* relation,

$$n \text{ bits} \Leftrightarrow 2^n \text{states} \quad (\text{exponential relation}) . \tag{I.2.38}$$

This implies that the converse relationship between information capacity and the number of accessible states is a *logarithmic* one: [9]

$$\#\text{ bits} = \log_2(\#\text{ states}) \quad (\text{logarithmic relation}) . \tag{I.2.39}$$

This relationship provides a precise and general quantitative definition of information that forms the very basis of information theory. The relation should remind you of the expression for the *entropy* $S = k \log W$ of a physical system, derived by Stefan Boltzmann, which links the entropy S as a state variable of a macroscopic system to the total number of distinct microscopic states W that correspond to that given macroscopic state, as we discussed in Chapter I.1 in connection with equation (I.1.62). So, entropy quantifies the microscopic diversity hidden in what we see as a single macroscopic state. In information theory, entropy is a measure for information capacity, the information that can be stored.

Entropy and probability. At this point it is interesting to refine this relation between available states and information by explicitly introducing the notion of *probability*. In the previous derivation we have tacitly assumed that given a single macroscopic state, the probability of finding the system in any of the corresponding microscopic states is uniform. With N states that would mean that $p_i = 1/N$ because the total probability should add up to $\Sigma_i^N p_i = 1$. In thermodynamics this distribution would correspond to a closed system at *fixed* (conserved) energy, and where one assumes the equipartion of energy.

[9] The information unit bit is linked to the logarithm base-2. If $S = \log_2 N$ this means that $2^S = N$. Thinking binary means that you reason in base-2. If I say a number is 21 in base-10, I make the statement that that number equals $21 = 1 \times 10^0 + 2 \times 10^1 = 1 + 20 = 21$. If I say a number is 21 in base-2, that statement makes no sense because the symbol '2' isn't there. To convert the number 21 in base-10 to base-2, I have to expand the number in powers of 2, so, $21 = 16+4+1 = 1 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 \Leftrightarrow 10101$. In base-10 the digits run from zero to 9, whereas in base-2 you only have 0 and 1. So the number 1011 in base two equals $1+2+8 = 11$ in base 10. This way, all integers can be uniquely encoded in any integer-based number system.

In general though, the probabilities will not be equal and one should introduce a probability distribution $\{p_i\}$ over the microscopic states, as we did in the section on statistical thermodynamics in Chapter I.1. There for a system in thermal equilibrium at temperature T, we introduced a probability p_i which was dependent on the energy ε_i of the microsystem labeled by 'i.' For that case we showed that the expression (I.1.69) for the thermodynamic entropy was first given by Gibbs, and now we see that it corresponds to an information entropy or information capacity (in bits) of the system given by the fundamental expression:

$$S = -\Sigma_i \, p_i \log_2 p_i \,. \qquad \text{(I.2.40)}$$

We mentioned already that the entropy of the system, as defined above, is equivalent to its information carrying capacity as it was defined by Claude Shannon. While working at Bell Labs he published in 1948 a groundbreaking paper on the transmission of information, that by many is considered to be the birth of information science. The important contribution from our point of view is firstly that he proved that it was the *unique* solution that satisfied some general constraints on information, and secondly that it applied in a general context that transcended its physical origins as thermodynamic entropy. So, that is where the term *information entropy* originated from.

Let us see what happens if we apply the formula to the two-spin situation where we have a set of four states which we denote as $\{11, 10, 01, 00\}$. We may turn on a weak magnetic field so that, say, the state 11 with both spins up is energetically preferred, for example leading to a distribution: $\{p_{11} = 1/2, p_{10} = p_{01} = p_{00} = 1/6\}$. Then the corresponding information capacity would be $S = \frac{1}{2}(1 + \log_2 6) = 1.79 \ \mathrm{bits}$, which is clearly smaller than the uniform case with all $p_i = 1/4$, yielding $S = 2 \ \mathrm{bits}$. The point I want to make here is that the uniform distribution is the maximally unbiased distribution, and it is that distribution which maximizes the information entropy, precisely because there is no additional constraint on, or in other words, 'additional knowledge' about, the system. Adding *a priori* knowledge reduces the information content, or the amount of surprise the outcome of measurements could provide. Constraints always reduce the number of allowed states for the system and therefore lower the entropy.

The Landauer principle. Talking about the relation between information and physical entropy it may be appropriate to briefly mention the principle proposed by Rolf Landauer in 1961, which is a particular formulation of the second law of thermodynamics which directly applies to information theory and computation. The principle expresses the fact that erasing information necessarily involves producing heat, thereby increasing the entropy. So, in other words, the principle governs the intimate relationship between information processing and the production of heat. This is of great importance, and it explains why large server parks tend to move up further north to colder environments. The heat produced by computers can certainly be reduced, but the improvements are bounded by the second law of thermodynamics.

We have illustrated the principle in Figure I.2.43. Consider a 'gas' consisting of a single atom in a symmetric container with volume $2V$ in contact with a heat bath. We imagine that the position of the particle acts as a memory with one bit of information, corresponding to whether the atom is on the left or on the right.

Erasing the information amounts to resetting the device to the 'reference' state $|1\rangle$ independent of the initial state, and therefore reinitializing the system rather than making a measurement. This can be done by first opening the diaphragm in the middle, then moving the piston from the right in, and finally closing the diaphragm and moving the piston back. In the first step the gas expands freely to twice the volume. The particle doesn't do any work, the energy is conserved, and therefore no heat will be absorbed from the reservoir. For that reason this is an irreversible free expansion process by which the entropy S of the gas

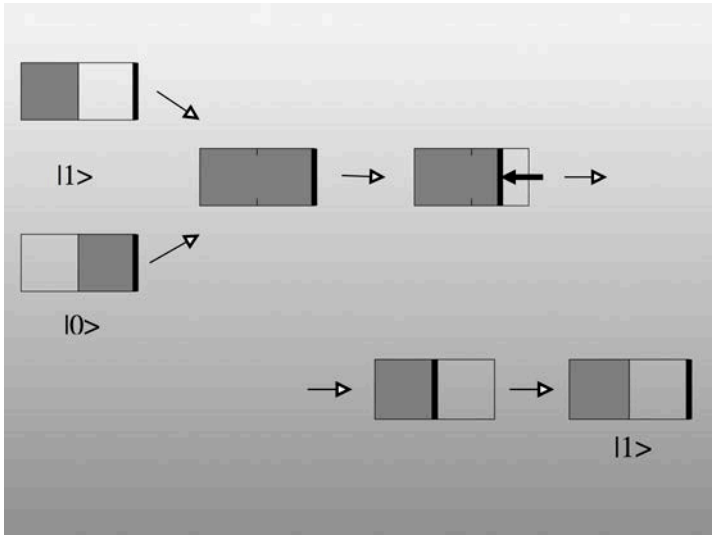


Figure I.2.43: *The Landauer principle.* An illustration of the Landauer principle using a simple 'thermodynamical system' consisting of a single particle in a vessel. See text for explanation of the successive steps.

increases by a factor $\mathrm{k}(\ln(2\mathrm{V}) - \ln\mathrm{V}) = \mathrm{k}\ln(2\mathrm{V}/\mathrm{V}) = \mathrm{k}\ln 2$. (The number of states the particle can be in is just the volume; the average velocity is conserved because of the contact with the thermal bath and will not contribute to the change in entropy). In the second part of the erasure procedure we bring the system back to a state which has the same entropy as the initial state. We do this through a quasi-static (i.e. reversible) isothermal process at temperature T. During the compression the entropy decreases by $\mathrm{k}\ \ln\ 2$. This change of entropy is nothing but the amount of heat delivered by the gas to the reservoir divided by the temperature, i.e. $\Delta S = \Delta\mathrm{Q}/\mathrm{T}$. Therefore the heat produced $\Delta\mathrm{Q}$ equals the net amount of work W that has been done in the cycle by moving the piston during the compression. The conclusion is that during the erasure of one bit of information the device had to produce at least $\Delta\mathrm{Q} = \mathrm{T}\Delta S = \mathrm{kT}\ \ln\ 2$ of heat. This argument shows that actually the heat computers generate is a necessary byproduct of them destroying information. It directly links the destruction of logical information with the thermodynamical generation of heat. This is a powerful result as it holds independent of the specific device one is talking about.

To summarize, you could say that 'forgetting' has its price (in heat). And that raises an interesting question about computation in general: can one avoid the heat by doing computation reversibly? The answer to this question was given by Charles Bennet in 1982, and is affirmative. However, reversible computation necessarily employs reversible gates only, but the familiar AND and OR gates (to be discussed shortly) are not reversible because they reduce a two-bit input to a one-bit output, producing at least $\mathrm{kT}\ln 2$ units of heat upon acting. A reversible computer doesn't pay the price of heat, but as all information has to be stored, the price of reversible computation is the requirement of ever-expanding memories! Not so cheap either.

Models of computation

> Computing is normally done by [a person] writing symbols on paper. [...] I assume that the calculation is carried out on one-dimensional paper, i.e., on a tape divided into squares. I shall also suppose that the number of symbols [...] is finite. [...] The behaviour of the computer at any moment is determined by the symbols which he is observing, and his 'state of mind.' [...] We may suppose [...] the number of states of mind which need to be taken into account is finite. ...the use of more complicated states of mind can be avoided by writing more symbols on the tape [...] Every [simple] operation consists of some change in the physical system consisting of the computer and his tape.
>
> Alan Turing,
> *On Computable Numbers with an Application to the Entscheidungsproblem*, Proc. Lond. Math. Soc. 2: 42. (1937)

Turing machines. Armed with a precise and operational definition of what information is, we should spend some time on computation or the processing of information. What are the basic underlying principles upon which the operation of all our computational devices is based?

We distinguish an input fed to a 'machine' that somehow processes that input leading to the desired result. To achieve this, the computer follows a sequence of instructions according to a certain procedure; an algorithm, or a program to produce the output. In a formal sense one could say that the device computes the output as a function of the input. As we have seen one can always present information in a binary way as a sequence of zeros and ones. So computers basically evaluate a function of the input, corresponding to the output. And a basic question concerning computation is to model this process in its full generality and determine what kind of functions can be calculated.

This is where the notion of a *Turing machine* comes in, which is a formal device satisfying certain specifications that can execute computations in the sense that it takes input and produces the desired output. It is not a machine in the ordinary sense but rather a fundamental model of computation. It does not address the question of the possible physical implementation of the models, of how to make them into real machines. It cannot care less whether you build it with rods and wheels, or like a fluid system with pipes and valves, or with Lego, or with elementary electronic semiconductor components called transistors.

Turing's starting point was in fact a rather natural and intuitive one based on the notion of an *effective computation*. A computation, procedure, or algorithm is called 'effective' if it satisfies the following criteria:

(i) it is specified in terms of a finite number of exact instructions,
(ii) if the instructions are carried out without errors, the desired result is obtained in a finite number of steps,

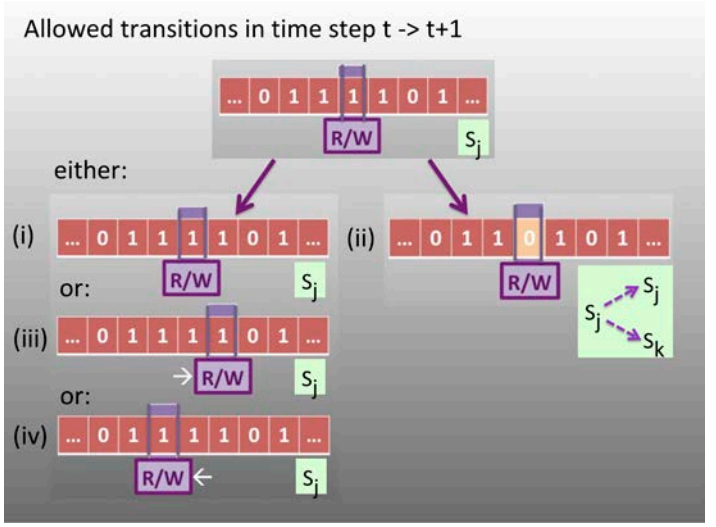


Figure I.2.44: *Turing machine transitions.* The four possible actions of the R/W head in a transition: (r) if it writes it cannot move in the same step, and the state may either change or not; (l) if it does not write it can move at most one step either to the left or to the right, but cannot change the state .

(iii) the instructions could in principle be carried out by a person only using using paper and pencil,
(iv) this person does not need any particular insight or ingenuity to carry out the instructions.

Note that there is no restriction on the amount of paper (memory), nor on the time it might take to perform the computation, apart from it being finite. The computation is 'effective' but not necessarily 'efficient.'

The Turing machine can in principle perform any such 'effective computation' and is defined as follows:

(i) it has a (half)infinite tape containing cells labeled by an integer p, each cell contains a symbol α taken from an alphabet $\mathcal{A}$. In the following we will just take the alphabet to be $\{0,1\}$, meaning that the tape is just a binary string which has a non-trivial input that starts on the left and may end with only zeros on the right.

(ii) a read/write head which is positioned at a given cell where it can read and (re)write the tape if instructed to do so. There are restrictions on what the head at any stage can do.
(iii) At any given time the machine is in some definite internal state S_j which is an element of some finite state space $\mathcal{F}_S$. The program or algorithm corresponds to a table that precisely specifies for every state what transition it has to make if the head reads either a zero or a one. This instruction specifies (a) what the head has to do, and (b) to what state the machine is supposed to go.
(iv) At the start of the computation, the input is the binary string on the tape. The head is located at the $p = 0$ cell and the machine is in the internal state S_0. The program halts if it reaches a final state (the output) where it finds no further executable instructions. So this is how a Turing machine computes a binary output function from some binary input.

From the fact that for any *effective computation* there is a Turing machine, one can prove the existence of a universal Turing machine that can perform all effective computations. This machine defines the set of *Turing-calculable functions.*

This rather intuitive definition of Turing-computability is the subject of the *Church-Turing thesis* which is central in the theory of computation. The Church-Turing thesis states that Turing computability is equivalent to the much more formal definition of computability based on *recursive functions* and *Abacus machines*. We are not going to dwell on these topics as they are really outside the scope of this book. The thesis cannot be proven as it links formal to intuitive notions. It is actually a hypothesis and all that can be said is that no counter example has been found so far.

At this point it is probably helpful to describe a basic version of the machine in some detail. In Figure I.2.44, we have the computer in some state S_j and we show the tape

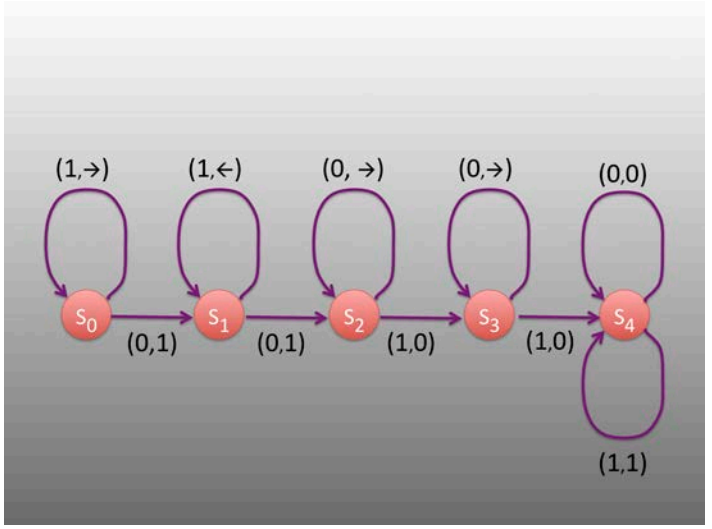


Figure I.2.45: *Turing machine state diagram.* The state diagram for the digital adder described in the text. In this program the machine goes through four states before it halts; in each state the move or write instructions on what to do if the head reads a 0 or a 1 are indicated.

with the R/W head at some position p. The program tells the head what to do but the possibilities are very restricted. There are only four possible transitions for the head/machine to execute:

(i) it stays at position p and does not change the entry with $S_j \to S_j$,
(ii) it stays at position p and does change the entry on the tape, in which case it also may or may not change the state of the system $S_j \to (S_j \text{ or } S_k)$,
(iii) it moves to the right $(p \to p + 1)$ with $S_j \to S_j$,
(iv) it moves to the left $(p \to p - 1)$ with $S_j \to S_j$.

The permitted transitions are schematically depicted in Figure I.2.44.

A Turing machine can also be represented by a finite state diagram. This diagram is a directed network where the nodes are the states S_j and the directed edges represent

the instructions. Instructions where the state does not change correspond to lines returning to the same state. The number of arrows leaving the node equals the number of symbols in the alphabet (in our case there are only two).

In Figure I.2.45 we have depicted the state diagram corresponding to a program that can add two positive integers m and n. We should think of the input as the two numbers in *unial coding* (this means that a number k is represented by a sequence of $k+1$ symbols 1) separated by a 0, with also zeros on the left and on the right. So the input sequence on the tape would look like:

$$000[11...11]_{m+1}0[11...11]_{n+1}000.$$

The head should then walk along the string of symbols starting from the most left 1, and then moves to the right till it hits the in-between 0, changing that 0 into a 1, so that the sequence then looks like:

$$000[11...11111...11]_{n+m+3}000.$$

Next the head should move to the left till it hits the first 0 on the left, then moves right again changing the first two 1 symbols into 0's. The result yields the required sequence representing the desired outcome.

$$000[11...11111...11]_{m+n+1}000.$$

You may verify that this sequence of steps is indeed performed by the machine depicted in Figure I.2.45, by following the sequence step by step.

We see that this simple problem already needs a quite complicated diagram. It is therefore more convenient to work in terms of logical gates, to which we now turn.

Logical gates. A computation is formally the calculation of a function f of many binary variables, so $f(a_1, a_2, \ldots a_n) = b$. The circuit for f should after entering an input of any set of a values return a binary number b. In practice one starts

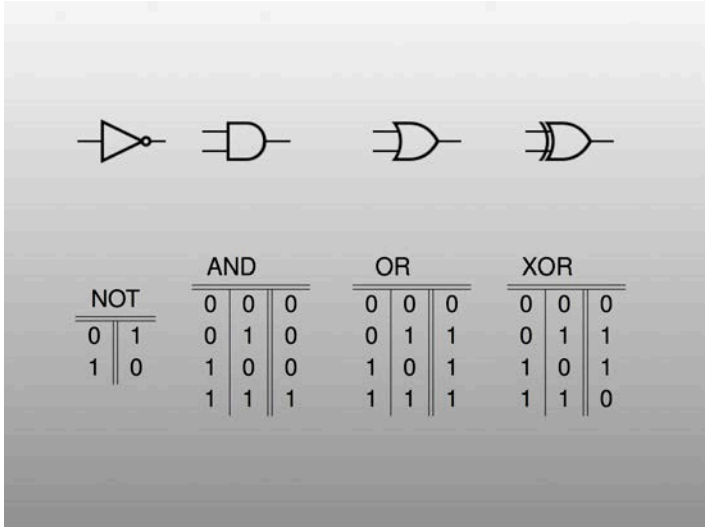


Figure I.2.46: *Logical gates.* The one-bit NOT gate, some two-bit gates, and their logical tables.

with a universal set of simple *logical gates* that compute certain basic functions. By combining many of those in specific parallel and serial arrangements, arbitrarily complicated functions can be composed. Diagrams with logical gates are simpler and more practical than going all the way back to the underlying Turing state diagrams.

The basic gates typically have only one- or two-bit inputs and a one-bit output, like:

(i) the NOT gate inverting the value of a single bit, meaning that if the bit contains a 1, then it is changed to a 0 and vice versa;
(ii) the OR and the AND gate. These are 2-bit gates, they are irreversible because they reduce the information of the 2-bit input to a 1-bit output.

One may prove that the set of these three gates is *universal*, in that they allow you to make machines to perform all the effective computations as defined by Turing. There are many other gates possible and these may be preferred depending on the problem one wants to solve,

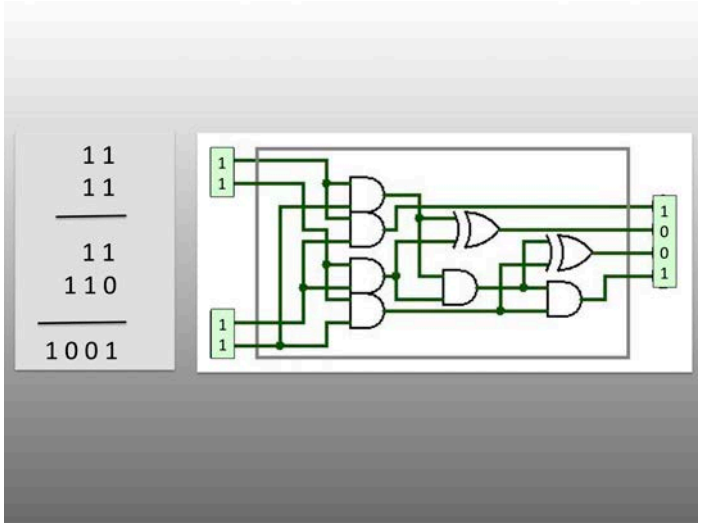


Figure I.2.47: *Multiplication.* A multiplication for two 2-bit numbers 'by hand' and the corresponding digital multiplier schematic composed of six AND gates and two XOR gates. The two lines of multiplication are performed in parallel, and the subsequent additions are sequential, so one therefore needs a total of $2n - 1 = 3$ time steps.

for example the *exclusive* OR gate also called XOR gate. It returns zero if both input bits are equal and one if they are different. These simple gates compute simple binary functions which can be represented in so-called *truth tables* where the possible input values (the arguments) are given on the left whereas the function value appears on the right. For the basic gates these tables are given explicitly in Figure I.2.46.

In Figure I.2.47 we demonstrate for example how to multiply two 2-bit numbers. We calculate $3 \times 3 = 9$, which in binary terms reads $11 \times 11 = 1001$. On the left we show how this is done by 'hand' with pencil and paper, and on the right how it is done by a logical device consisting of some AND and XOR gates. Using the truth tables it is quite straightforward to follow the lines and put the bit-values on them, and convince oneself that it indeed works.

Going quantum

> Until recently, most people thought of quantum mechanics in terms of the uncertainty principle and unavoidable limitations on measurement. . . The appreciation of the positive application of quantum effects to information processing grew slowly.
>
> Nicolas Gisin

Once we have come to appreciate the basic fact that information capacity is directly related to the 'number' of available states of a system, it is immediately clear that if we are to descend to the level of quantum mechanics, we have to think in terms of quantum states. As we will see, quantum states are quantessentially different from their classical precursors, and therefore we should be prepared to go back to the drawing board and define from scratch what we mean by information. The space of states has a completely different structure indeed, and that forced the scientists to start developing what is nowadays called *quantum information theory*.

It is in that way that a turning point in our understanding of what matter really is on the microscopic level induced a radical change in our basic notion of information. It was the eminent physicist Richard Feynman who maybe for the first time pointed out some of the basic principles in a well-known paper entitled *There is plenty of room at the bottom*. The change did not just affect the abstract, software side of information theory, but also the hardware side. The crucial challenge is nowadays to develop new types of quantum technology that allow us to store and manipulate quantum information. Without exaggeration one may say that this constitutes a new holy grail for experimental physics and engineering.

There are basically two reasons why information will go quantum. The first is that information science has to confront quantum physics at some point because of Moore's

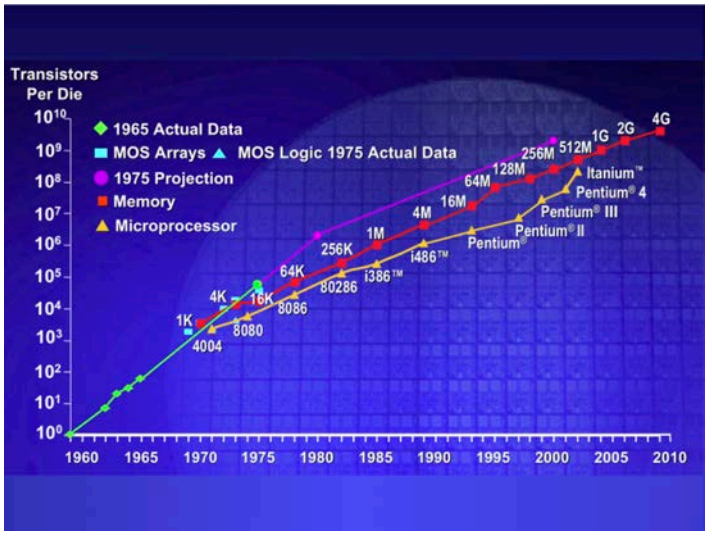


Figure I.2.48: *Moore's law.* This law states the astonishing fact that over the last half a century the power of computing has doubled every 18 months. The continuous downscaling of the basic components forces us to enter the gates of quantum domain. (Source: High Tech Forum)

Figure I.2.49: *RSA-2048.* RSA-2048 is a number with 2048 binary and 617 decimal digits. The factorization has not been found yet.

law. The second is that scientists who looked more thoroughly into the equations governing quantum information made the astounding discovery that for a number of tasks the quantum computer is extremely more powerful than its classical digital counterpart.

Moore's law. This is an empirical 'law' which refers to the spectacular fact that our computational power over the last half a century has increased at an incredible rate: on average it has doubled every 18 months, as you can see in Figure I.2.48. This implies that it has been growing exponentially for more than half a century! We are now at a stage where a single active component of an integrated digital circuit has a size of about 10 nanometer, very small indeed. Once you realize that atoms are of the size of a nanometer, it is clear that Moore's law has to break down if we don't succeed in entering the quantum domain. In other words the continued scaling down in the size of the hardware components forces us to enter the quantum world one way or another!

A tough problem: integer factorization. But going quantum also means that we turn something that at first sight looks like a crisis into a tremendous opportunity. Quantum mechanics is so fundamentally different, that it would allow for a quantum computer to solve problems that would be intractable on our classical digital computers.
A famous example is the factoring problem: I give you a very large integer N of n digits which I tell you can be written in a unique way as the product of two other integers M_0 and M_1 . I don't tell you what they are, but instead ask you to find M_0 and M_1 . This turns out to be an extremely hard problem not only for people but also for very, very big computers. Hard in the sense of time needed to find the answer. Numbers of this type, that can be factorized into two prime factors are called RSA numbers and they have important applications in cryptography.

That may surprise you but let us get a rough idea of why this is so.

A simple way to find the divisors is the method of 'trial division' which goes back to the medieval Italian mathematician Fibonaci. To know whether a number N has a divisor M you start with N and keep subtracting M until after k steps you get a number smaller than M, if that number happens to be zero then M is a divisor of N. You start doing this by choosing $M = 2$ and that takes care of all even divisors. Clearly the next number M we have to check for would be the next prime number but that requires that the list of primes is known. To get a rough estimate what we can do is to check divisibility for all odd divisors. One additional observation that simplifies the search is the fact that if the two prime factors are unequal then one will be larger than $\sqrt{N}$ and the other smaller. We thus have to check the divisor property only up to $\sqrt{N}$. Knowing that apart from the number 2 all prime numbers are odd we have to only search for odd divisors, which leads to a further reductions. An estimate for the maximum number of simple subtractions P^* in such a worst case scheme would give:

$$P^*(N) = \sqrt{N}\Big(\frac{1}{2} + \sum_{k=1}^{\sqrt{N}} \big(\frac{1}{2k+1}\big)\Big) \tag{I.2.41a}$$

$$\simeq \frac{\sqrt{N}}{2}\Big(1 + \int_1^{\sqrt{N}+1} \big(\frac{1}{x+\frac{1}{2}}\big)\,dx\Big)$$

$$\simeq \big(\frac{n}{2}\ln 2\big)\,2^{n/2}\,. \tag{I.2.41b}$$

In the last line we have assumed the number N to be a n-bit number, $N \simeq 2^n$, and kept only the leading term in n. The key conclusion we draw from this rough estimate that the core time needed to factorize an n-bit RSA number grows exponentially with n. It is no surprise then that children find factorizing to be much harder than multiplication, and that is why in the pre-calculator-era they had to learn the multiplication tables (which are also factorization tables) from 1 to 20 by heart, like it concerned the first few couplets of a universal human anthem! And with computers we do now the same thing, reading values from tables, whether they like it or not. A realistic example of such a gigantic number is RSA-2048 shown in Figure I.2.49, having 617 digital or 2048 binary digits. It is a public challenge to factorize it into two primes, and if you meet the challenge you get US$ 200.000 – unfortunately the number of dollars does not come near N, nevertheless making it worth to give it a try! But wait is that true? We just calculated that the amount of processor time would typically be $t^*(2048) \simeq P^*(n = 2048) \times (10^{-10}\,sec) > 10^{300}\,yr$. this is a clear warning that you have to come up with a rather smart idea.

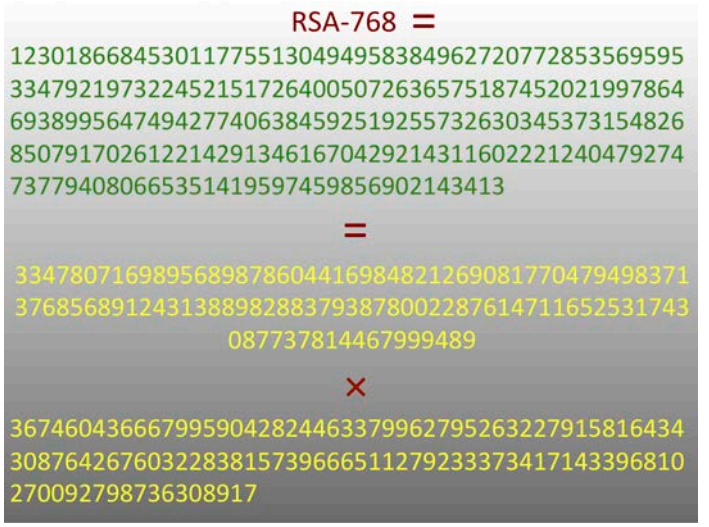


Figure I.2.50: *RSA-768.* RSA-768 is a number with 768 binary and 232 decimal digits. The factorization given below was obtained through a heroic effort by an international collective of experts. It would have taken a powerful super-computer some 2000 years, but they managed to do it in just two years.

An example of an integer number that – in a heroic effort by an impressive international collective of computer experts and mathematicians, using a tremendous amount of algorithmic ingenuity and digital power – has been successfully factorized in its two prime factors, is called *RSA-768* with 768 binary or 232 decimal digits. The result is displayed in Figure I.2.50.

Quantum factorization. We concluded that with a classical computer the typical time it takes to factor N in its prime factors grows exponentially with its size n, but the American applied-mathematician Peter Shor proved in 1994 that with a quantum computer the job can be done in *polynomial* time. We will discuss the (quantum) algorithm he constructed in more detail towards the end of Chapter II.4 in Volume II.

The factorization problem is in a strange way asymmetric: finding the integers M_0 and M_1 is kind of exponentially hard, but if you give me those integers, you and I can simply check wether you are right by just multiplying them using a large calculator, in a time of order $t \simeq n$. Factorization is one of the main tools in cryptography, so it is not just a matter of academic interest. It is of prime interest to all those who are concerned about security and safe transactions via the internet, like banks (and their clients), medical services, intelligence agencies and twittering celebrities. In fact, with today's world in a severe state of *cybernation*, all of us are highly dependent on a secure internet!

To see the huge importance of exponential vs. polynomial scaling, suppose an elementary computational step takes Δt seconds. If the number of steps increases exponentially, factorizing a number with n-bit will take $\Delta t\, 2^{an}$ seconds, where a is a constant that depends on the details of the algorithm. We have depicted some of the different computation time behaviors in Figure I.2.51. The take-home message there is the huge qualitative disparity between polynomial and exponential behavior that becomes manifest for large n.

For example, if $\Delta t = 10^{-6}$ and $a = 0.1$, factoring a number with $n = 1,000$ binary digits would roughly take 10^{37} seconds, which is much, much longer than the lifetime of the universe (which is a mere 4.6×10^{17} seconds). In contrast, if the number of steps scales as the third power of the number of digits, the same computation takes $a' \Delta t\, n^3$ seconds, which with $a' = 10^{-2}$ is 10^4 seconds or a little under three hours. Of course the constants a, a' and Δt are implementation dependent, but because of the dramatic difference between exponential versus polynomial scaling for sufficiently large n, there is always a huge qualitative gap in speed that cannot be compensated for by adding more pieces of conventional hardware.

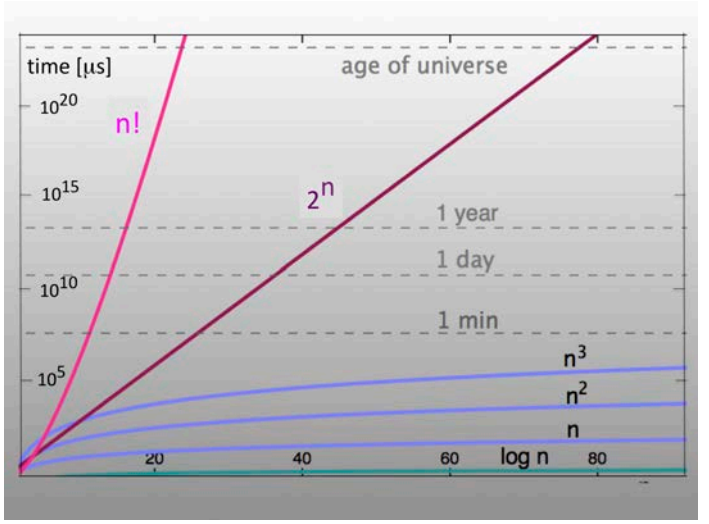


Figure I.2.51: *Computational complexity.* The classes P and NP refer to the growth of time needed to solve a problem of size n. Problems in P can be solved in polynomial time ($t \sim n^\alpha$ for some number α) and problems in NP cannot. These might grow exponentially ($\sim 2^n$) or super-exponential (like the factorial $\sim n!$.) (Source: C. Moore, SFI)

I should add that for the factoring problem as such, the situation is in fact more subtle: at present the best available classical algorithm does significantly better than exponential, it would require $O\big(\exp(n^{1/3}\log^{2/3} n)\big)$ operations, whereas an available quantum algorithm provided by Shor needs $O\big(n^2 \log(n) \log(\log n)\big)$ operations. To give you an impression we give a log-linear plot of the two factorization times in Figure I.2.52, and you can see that the behavior for large n is qualitatively drastically different with slopes tending to 1/3 (classical) and zero (quantum).

Factorization is only one of several problems that could

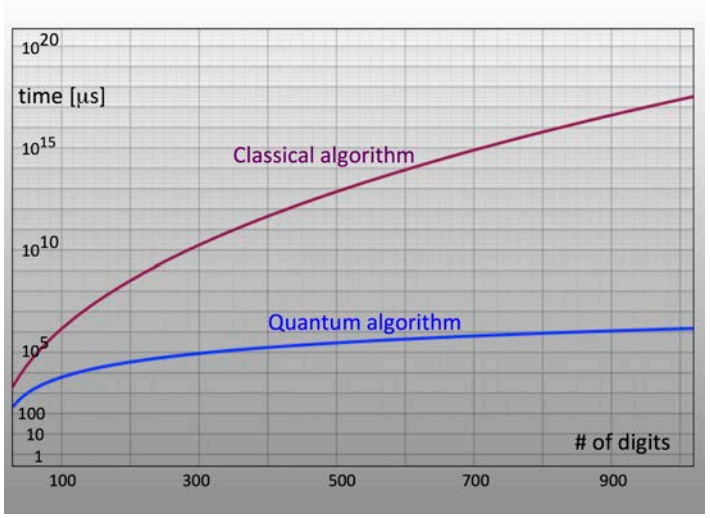


Figure I.2.52: *Factorization algorithms.* A log-linear plot of the estimated time it takes to factor an n digit number with the best available classical and quantum algorithms mentioned in the text.

potentially benefit from quantum computing. The implications of quantum information even go beyond quantum computing, and include diverse applications such as quantum cryptography and quantum communication, which by the way is intrinsically secure.

The quantum leap to such mind-boggling speed-ups arises from two main sources. Firstly from the intrinsically parallel nature of quantum mechanics, which in turn is a consequence of a quantessential feature called the *linear superposition principle*. This parallelism basically derives from the fact that state vectors have many components, and a quantum interaction or operation or gate affects all components simultaneously. Secondly from the existence of so-called *entangled states* that are unique to quantum theory. Particles that are in an entangled state can be correlated in a way which is not possible in classical physics. We will talk in quite some detail about these quantessential notions in Volume II. The actual workings of quantum theory were apparently sufficiently subtle that it took many decades after the discovery of quantum mechanics before anyone realized that its computational potential was fundamentally different and quite powerful indeed. The huge interest in quantum information and computation in recent years has caused a thorough re-examination of the concept of information contained in physical systems, spawning the field that is referred to as 'quantum informatics.'

Computational complexity. One of the deeper issues in the theory of computation is to try and quantify what we mean by *computational complexity*. Roughly speaking a measure of the complexity of a problem would be the time it takes to solve the problem on a computer running an optimal program (algorithm) for that problem. The time it takes to multiply two n-digit numbers on a computer for example would naively grow quadratically with their size n, because you have to do of the order of n^2 basic multiplications (plus order n additions). You can gain a factor n by parallelizing the algorithm: the multiplications giving the n 'rows' in the standard multiplication chart can be done in parallel, and the subsequent additions have to be done sequentially, as indicated in Figure I.2.47. The classification of complexity is now linked to the functional dependence of the computation time on n.

There is a crucial distinction to be made here. Firstly, there are problems that can be solved in polynomial time, meaning that time is bounded by some simple power law $t \leq n^k$. Such a problem is by definition in the 'polynomial' class P, but one believes that there are many problems that do not belong to P and they belong to a larger set containing P as a subset denoted by NP. Note that NP does not just mean 'not polynomial.' The set NP contains problems of the 'find-the-needle-in-a-haystack' type. These are hard to solve because you basically have to do an exhaustive search of the whole stack and that takes a hell of a lot of time. The distinguishing property for NP is that once you have found an answer it is straightforward to check that your answer is right or wrong. Easy, because a needle is a

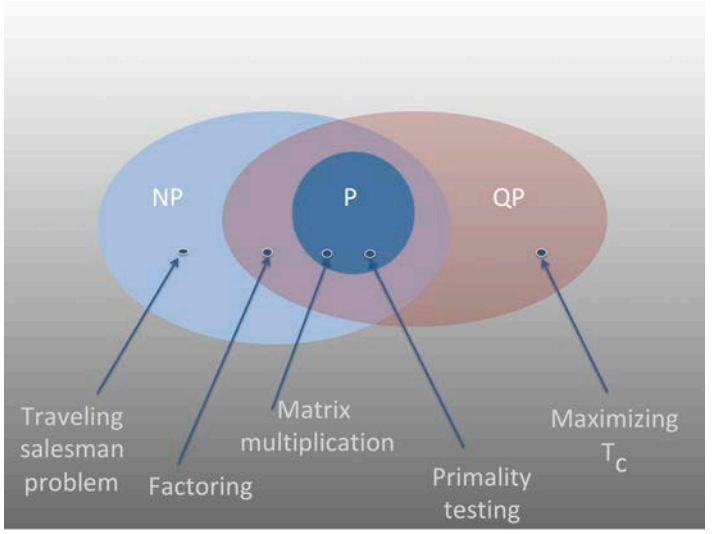


Figure I.2.53: *Complexity classes.* Hypothetical hierarchy of computational complexity classes and some standard problems belonging to them. Note that the integer factorization and graph isomorphism problem are classically believed to be not in P but in NP while in quantum informatics they belong to QP. (Source: M. Freedman et al.)

needle, isn't it? The formal statement is that the answer to an NP problem can be checked in polynomial time.

The hardest problems in NP are called NP-*complete*. The NP complete problems are in an abstract way equivalent, meaning that they can be mapped onto each other in a one-to-one way. If you solve one, you have solved all of them. Integer factoring is believed to be in NP but not in P. Furthermore the problem is *not* considered to be NP-complete; it is believed to belong to an intermediate class. The complexity of complexity theory is that we do not a priori know that a super smart algorithm does not exist to factor large integers into their prime factors in polynomial time, but that we just have not been able to find the algorithm yet, nor have we found a formal proof that such an algorithm does not exist. We find ourselves in a serious *catch-22* situation. Therefore one likes to say that certain problems are 'believed' to be NP-*complete*.

P versus NP. Indeed, the million dollar question really is whether NP in the end is not just *equal* to P! Here we just have to wait for some real or arificial computer genius to strike. That question by the way is considered to be so fundamental, that it appears on the illustrious list of seven Millennium problems of the Clay Institute for Mathematics in the US, which were announced at a meeting in Paris, held on May 24, 2000 at the Collège de France. Just solve it and they will pay you that million dollars!

Clearly the advent of quantum information theory calls for a new complexity classification scheme, with new categories denoted as QP and QNP. And therefore the complexity analysis becomes even more intricate. Whereas factorization is believed to be classically NP it is in quantum QP as we have indicated in Figure I.2.53. Nevertheless, as things stand now, there is still a remote but dramatic possibility that the content of this complexity picture in the end collapses to a single point!

We will return to what a *qubit*, the fundamental building block of a quantum computer, exactly is, as well as to the basics of quantum communication in Part II of the book. Quantum computation as a branch of science nowadays involves sophisticated and highly specialized subfields of experimental physics which are beyond the scope of this introductory book. We want to restrict ourselves to the quantessence after all. One quantessential conclusion we want to draw here is that information will go quantum not too long from now. Or, to quote Nelson Mandela: 'It's always impossible until it's done.'

Quantum physics: the laws of matter

> [The homeland] looked strange to us returned soldiers... The civilians talked a foreign language. I found serious conversation with my parents all but impossible.
>
> *Robert Graves, Goodbye to All That.*

Understanding the deep structure of matter has led to a new conceptual basis for all of physics. A basis that governs the laws of new fundamental particles and force fields but also of new phases of condensed matter, of chemistry and finally the laws of quantum information.

Surprisingly, this section is the shortest of this chapter. The reason is simply that we still have a whole book in front of us on the subject. Quantum theory has the names of many great scientists associated with it, and not just because of the saying that success always has many parents. Roughly speaking one distinguishes three generations of quantum physicists. The first generation consists of people like Max Planck who coined the idea that energy of heat radiation be quantized, Albert Einstein who, following Planck, postulated the existence of a particle of light, which he called a photon and explained the photoelectric effect using this new particle, and finally, Niels Bohr, the great Danish physicist whose model for the atom proved it to be a tremendous breakthrough. A second generation consists of great names like Erwin Schrödinger, Werner Heisenberg, Paul Dirac and others, who managed to give a mathematical foundation for the theory and derive its fundamental equations. Many other luminaries like Wolfgang Pauli, Max Born, Enrico Fermi and John von Neumann greatly enhanced our understanding and interpretation of the theory (see Table B.1 on page 645 of Part III).

After the Second World War a third generation took the stage, with the development of quantum field theory as the most outstanding fundamental contribution. Great physicists like Richard Feynman, Julian Schwinger and Sin-Itiro Tomonaga completed quantum electrodynamics shortly after the war, and during the sixties and seventies a long list of distinguished scientists constructed the Standard Model of elementary particles and fundamental forces (see Table B.3 on page 647 of Part III).

Parallel to these developments many new research directions opened up such as quantum chemistry, quantum condensed matter theory, quantum material science and quantum optics (see Table B.2 on page 646 of Part III). We would also like to mention the fundamental progress in our theoretical understanding of quantum principles that these three generations and generations after them have left us with. This book is of course completely devoted to these matters and we will discuss what the central ideas of quantum theory are and how counter-intuitive and therefore unbelievable these ideas must have appeared at the time of their inception. You might experience some of that same uneasiness as you read along. As a matter of fact quantum physicists all around the globe have acquainted themselves with the theory to such a degree that most of them have developed some kind of 'quantum intuition.' And yet, in spite of that they are still regularly taken by surprise with what nature is telling them.

The development of quantum theory is one of the most astonishing achievements of twentieth century science to which a large number of gifted characters have contributed in the period of time encompassing the two world wars. It paved the way for a multitude of technological advances and even now we feel that the era of quantum technologies has only just started. This is exemplified by the promising developments where quantessential principles are exploited to create a totally new type of information science, involving quantum computing, quantum teleportation and quantum cryptography. Such is the power of truly new fundamental insights in the workings of nature: what at first appears as pastimes for absent minded eggheads, ends up as core ingredients of radical innovations and new technologies. Innovations that have offered new options for society, and often have deeply affected the human condition.

This book is quite voluminous, but that should not surprise you once you realize that – as is in full display in the tables at the end of the book – so many Nobel prizes have been awarded in this incredibly prolific field of science.

Further reading.
On relativity:

- *Very Special Relativity: An Illustrated Guide*
  S. Bais
  Harvard University Press (2005)
- *Exploring Black Holes: Introduction to General Relativity*
  E.F. Taylor and J.A. Wheeler
  Addison Wesley (2000)
- *General Relativity*
  R.M. Wald
  University of Chicago Press(2010)
- *Gravity: An Introduction to Einstein's General Relativity*
  J.B. Hartle
  Cambridge University Press (2021)

On the physics of geometry:

- *Flatland: a Romance of Many Dimensions*
  E.A. Abbott
  Penguin Group (2020)
- *The Geometry of Physics: An Introduction*
  T. Frankel
  Cambridge University Press (2011)

On the Physics of Information:

- *Introduction to the Theory of Computation*
  Sipser
  Cengage India (2014)
- *The physics of information*
  F.A. Bais and D. Farmer
  Chapter in *Philosophy of Information*
  P. Adriaans and J. van Benthem (Eds)
  Elsevier Publishers (2008)

Chapter I.3

Universal constants, scales and units

Is man the measure of all things?

Physicists have come to appreciate the existence of certain universal constants of nature like the velocity of light, Newton's constant, the elementary charge, Planck's constant etc. These are numbers that cannot be calculated from first principles. They have to be obtained from measurements and their values set the scales that characterize our universe. First we show how these constants can be used to define a complete and consistent system of units. In the second section, we take a step back and ask whether these constants are really universal, or just the parameters that appear in our theories and therefore only reflect the present state of science. In the third section, we play around with these constants to explore to what extent these natural scales mark the domains of validity of particular theories. We conclude by describing the Planck system of 'natural' units and discuss its interpretation. Indeed, the arguments presented in this chapter suggest that man is not the measure of all things, rather the arguments constitute a modest plea to bid farewell to anthropocentrism.

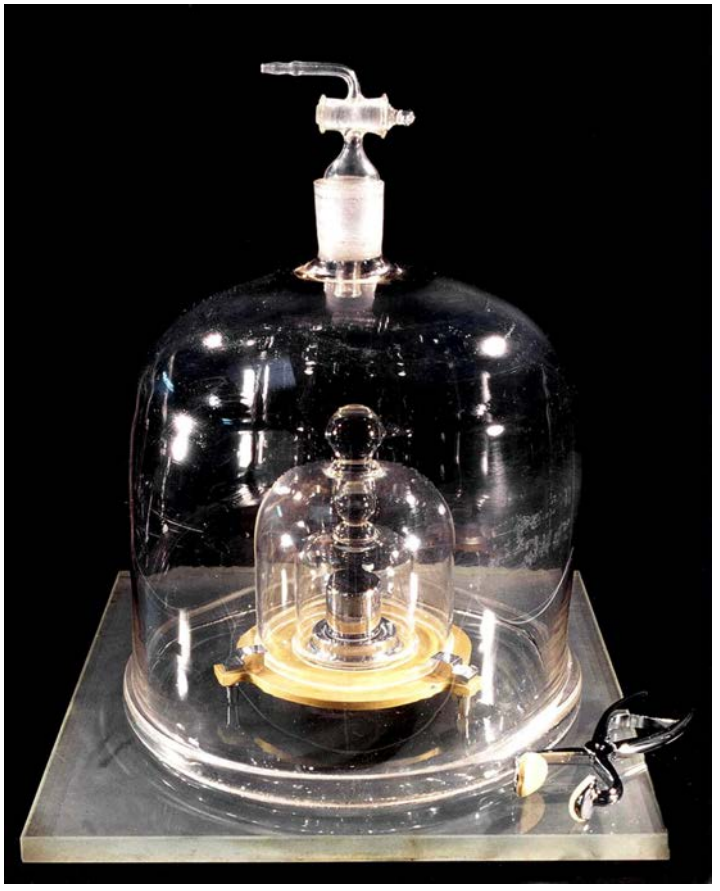


Figure I.3.1: *The international prototype of the kilogram.* Up to 2019, this was the standard of mass, kept under three glass bells in the *Bureau International des Poids et Mesures* in Paris. (Source: Wikimedia)

Isn't it a pity that we have lost many of those good old home and kitchen units, such as the thumb, the ell, or the foot, the knifepoint, the stone, the cloud, the crate, the walking hour, or horse power? The 'foot' is an example of where man was taken as the measure of all things; in fact

it was in the middle ages around 1100 that King Henry I decreed that *his* foot would be the unit of length. A nice illustration of the amusing fact that even only *one* man could be the measure of all things, with the clear disadvantage that that unit undoubtedly changed over time and furthermore one may assume that he took this 'standard foot' with him in his grave. Since the era of the Enlightenment we have been 'decimalized,' as the powers of ten are naturally built into our common number system and our metric unit system, and they are now by far superior, not least because they are pretty much shared globally. However as you know there are many remnants that don't fit in. I am not just talking of astronomers or atomic physicists claiming that they only live 10^{-13} *parsecs* or 10^{13} *angströms* respectively from their work, because such jargon is presumably rather a measure of their 'professional deformation,' or should I say devotion? A brief history of time may clarify what I mean.

On time

It's about time. In spite of the globally accepted metric supremacy, there remains ample room for exceptions. Think of our units of time for example. As you know, these are mostly dictated by the dynamics of our solar system, with the *year* that refers to the earth's rotation around the sun, while the *month* is set by the moon's rotation around the earth and the *day* is fixed by our rotation around the axis of the earth. In fact the system of time divisions was primarily inspired by the geometry of the circle, which has 360 degrees, approximately one degree per day. The circle exactly encloses six adjacent equilateral triangles with all angles equaling 60 degrees, and when you cut this partition in half – which one can do with only ruler and protractor – you would account for the division of a year in twelve months. The solar system's periodic motions serve as a celestial clock, with the almost natural choice of 24 hours to the day. It is better to think of twelve hours for the day and twelve for the night, which is a division believed to go back to the Egyptians who did their arithmetic in base 12. From the hour down, the minute and the second are then counted in the base-60 numbering[1], and below the second we talk milli- and nanoseconds and we unanimously convert to base-10 numbering. At the opposite end of the scale we think also in powers of ten centuries and millennia. So, indeed, our common time units are quite archaic and convoluted.

Unifying the incommensurate. The numbers given to us by Mother Nature are far from accurate because they may vary. Moreover, they inhibit implementing the geometric precision we just alluded to, because there is no physical reason why the units of year, month and day should have anything to do with each other as they refer to entirely different dynamics which are almost completely decoupled. And that's of course why the year is approximately $365.2422...$ days. To put it in perspective, it is like decreeing that from now on there are approximately $9.893...$ cents to the dime and 9.734 dimes to the dollar! Such incommensurate units would lead to a lot of problems at the check out, I am sure!

To arrive at an orderly bookkeeping of time it took nothing less than a pope – Gregory XIII to be precise – to decree in 1582, much like a well-trained engineer, that we should make successive approximations. First we put 365 days in the year, but to make up for the other decimals we add one day – let's pick the 29th of February – every four years, and call that a *leap year*. That brings us up to 365.25 days per year on average. Now the next step in our approximation is made by skipping one *leap year* at the turn of the century, which brings the leap day contribution down by a factor $1/25$ so we drive at 365.24 days per year on average. In the next step, we don't skip every 400 years

[1] The base-60 or *sexagesimal* number system goes back to the Babylonians as far as about 3100 BC. They later even introduced a positional notation marking for empty places (like our zeros) to keep track of additional powers of 60.

which gives us a score of 365.2425.The subsequent corrections are accounted for in a rather ad hoc manner by the introduction of what are called *leap seconds*.

We see the disparity and feel the tension between nature's innate rhythm and the strictly rational recipes we would like to impose. It reminds us of that funny story of a governor of a southern state in the US, who thought he could render his community a great service by decreeing that the number π from then on would be set equal to 3 in order to simplify life! But as the number π is defined as the ratio of the circumference to the diameter of a circle, there is not much room for decreeing anything about it. With the millennium debacle still fresh in our minds, when lots of computer software went haywire because of hardwired calendar settings which couldn't handle the number 2000, we may have to anticipate future troubles simply because the trivial accounting of the Gregorian calendar has not been implemented correctly.

It is amusing to learn that the decimal metric system, which goes back to the French Revolution, was also originally intended to cover the measurement of time. In 1793 apparently the French Republican Calendar was introduced, with weeks of 10 days, lasting 10 hours, with 100 minutes to the hour, and 100 seconds in one minute. This caused massive protests, not in the least by the church authorities, who felt they were losing influence and didn't want to reshuffle their Holy days, which were shared anchor points for people's sense of time. It was only in 1805 that Napoleon decided to abandon the system.

The system of time units is, like our DNA, the outcome of a contingent sequence of improvements that for the case at hand co-evolved with us humans. Our common units of time unmistakably reflect the subsequent stages of human scientific awareness and technological advancement.

Reinventing the meter

An optimal system of units should be complete and consistent, but also precise. This implies that the most advanced measurement of the universal constants of nature, or combinations thereof, have to be used to define units. According to the *Système International (SI) of units*, it distinguishes 7 *base units* and more than twenty *derived units*. The 7 (independent) *base units* are: the *second* (time), the *meter* (length), the *kilogram* (mass), the *ampère* (current), the *kelvin* (temperature) the *mole* (amount of substance) and the *candela* (luminosity).

The measurements by which these units have to be defined should not only be precise, but should also be relatively easy to reproduce, so as to make it easier to share the system of units in a practical way. These criteria are ever more relevant, as many of our daily activities depend on a great precision of measurement that makes our devices work, think for example of using the Global Positioning System (GPS). These criteria also make it mandatory that the system of units has to be upgraded from time to time so as to take advantage of the newest scientific and technological advances, not unlike the operating systems of our computers.

Let us return to our brief history of time, and see what happened to the definition of the *second* as a unit of time in the course of time. We started with time units inspired by the heavenly mechanics and the observations thereof. It may surprise you, but indeed, up to 1960 the second was defined as 'the fraction $1/86400$ of the mean solar day.' The exact definition of 'mean solar day' was left to astronomers. Apart from the fact that the rotation of the earth has irregularities, the measure itself was ad hoc. In 1967, it was finally switched from an astronomical to an atomic time standard as it is both far more precise and much easier to reproduce.

When the Saints go marching in... Given the plain fact that the human length is of the order of a meter, their weight is in the range of kilograms, their hart ticks at the rate of seconds we should not be surprised that we have ended up with something nice like the metric SI *(Systeme International d'unités)* – or MKS (Meter-kilogram-second) system as the measure of measures. And with it come the *prefixes*, the formal powers of ten, from picoseconds to terabytes and beyond.This metric thinking suggests that scientists have lifted their quantitative thinking entirely to the rational norm.
But alas, it is exactly in their ranks that irrational alternatives flourish. Extensive use is made of *derived units* that pay tribute to their ancestors and perhaps – who knows – one day to themselves. Experts thus actively employ the *Newton, Joule, Pascal, Coulomb, Watt, Farad, Ångström, Tesla, Gray, Henry, Fermi, Ohm, Siemens, Weber, Hertz, Oersted, Becquerel, Rydberg, Curie, Fahrenheit, Röntgen, Stokes, Millikan, Gray, Sievert,* and whatnot. What's in a name, you may wonder. However, note that *we should have typeset these names in lower case*, to avoid any suggestion that they might refer to individuals. After all, the force of 3 Newtons is quite something else than 3 newton. If only we could have 3 Newtons! It reminds me of the disclaimers made in the preface of some classic novels: 'all similarities with persons alive or dead are purely accidental.'
Count your blessings though: in the nineteenth century, just to communicate about temperatures, one had to convert between a rich variety of what I would like to call *tribal scales*. Not only the familiar degrees *Fahrenheit, Celsius* and *Kelvin*, but also degrees *Réaumur, Rømer, Rankine* and *Wedgewood*! Fortunately there is only one nature, meaning that whatever units you happen to invent, they always can be converted to more sensible ones. So, referring to obscure units is more a matter of name-dropping highly-esteemed colleagues, than using double standards. □

Today the *second* is defined as:

> the duration of *exactly* 9192631770 periods of the radiation corresponding to the transition between two hyperfine levels of the ground state of the Caesium-133 atom.

This we may write as an exact defining equation:

$$\nu_{\mathrm{Cs}} \equiv 9,192,631,770\ s^{-1}\,. \tag{I.3.1}$$

This definition of the second refers to the frequency associated with the radiation that is transmitted if the Caesium atom makes the transition between two well-defined quantized energy (hyperfine) sub-levels. You could say that a Caesium clock gives about 9.2 billion ticks a second. That quantity can be measured with great precision, meaning that if you compare the outcomes of a great many carefully performed measurements, the spread of outcomes will be extremely small. In other words it is the spread of these measurements which determine the number of significant (reliable) digits. By defining the second as a *fixed* number times a physical observable, the number of significant digits in the definition of the unit equals that of the best possible measurements. The central point here is that the units inherit the precision of the measurements and they therefore necessarily co-evolve with the state of the art in experimental physics, without the need to redefine the units all the time.

You may not be surprised to hear that at present physicists are in the process of developing devices which will allow us to define the unit of time by a factor 100,000 times more

precise', by using so-called *femto second lasers*, that deliver tiny pulses about 10^{15} per second. This technique uses a so-called *frequency comb*, produced by a pair of frequency locked optical lasers. It is a quantum optical device for which the Nobel prize was awarded in 2005 to the American physicist John Hall and his German colleague Theodor Hänsch. Indeed the definition of time is getting outdated all the time and a switch to the new quantum optical standard is to be expected in ten years' time.

In quantum theory many observable quantities like energy levels, currents, fluxes, charges and so on turn out to be quantized, meaning that they only can take on discrete values, exactly equal to integer multiples of certain combinations of universal constants. This 'quantization' property allows them to be measured with extreme precision and that makes them particularly suitable for defining units. We should devise definitions for a set of base units linked to the universal constants of nature so that we can measure the best, and then use those to define the other derived units.

Also in that vein the unit of length, the *meter*, was redefined in 1983 as:

> the distance traveled by light in vacuum in *exactly* $1/299792458$ of a second.

Another way to say it would be to state that,

$$\mathrm{c} \equiv 299792458\ \mathrm{m/s}\,, \tag{I.3.2}$$

again exactly, no decimals to be added! This definition together with the definition of the second then *defines* the meter. We need no longer refer to the *International Prototype Meter* kept at the *Bureau International des Poids et Measures* in Paris, as the distance between two marks on a Platinum-Iridium bar that was kept at the freezing temperature of water.

Now, it may come as a surprise to you that the definition of the *kilogram* as the unit of mass was up to 2019 linked to an artefact, the *International platinum-iridium kilogram* kept at the aforementioned *Bureau* in Paris, and shown in Figure I.3.1. It comes across as indeed somewhat archaic, and fortunately this artefact has been replaced by a more adequate and operational definition involving Planck's constant, again referring to precise measurements of quantum behavior.

The definition of ampère also used to be somewhat cumbersome and hard to implement. It was defined as:

> the constant current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross-section, and placed 1 meter apart in vacuum, would produce between these conductors a force equal to 2×10^{-7} newton per meter of length.

Imagine entering the store and asking for two infinite wires of zero cross section: 'Oh yes, Sir, uh, let me see, oh no, it's not in the catalogue. I am really sorry Sir. And by the way, Sir, may I ask also you to be so kind as to leave my store immediately please.'

As to the notion of temperature, the definitions were linked to phase transitions in matter systems, as for example the Celsius degree which was defined as 1/100 of the temperature difference between the boiling and freezing temperatures of water under 'normal' conditions. Since 1954, the kelvin has been defined as exactly equal to the fraction 1/273.16 of the thermodynamic temperature of the triple point of water, which is the point at which water, ice and water vapor co-exist in equilibrium. That is a very useful definition because for water at a specific pressure, the triple point always occurs at exactly a temperature of 273.16 K. Yet also there it was agreed to couple the definition with a universal constant – the Boltzmann constant k – which links energy and temperature according to the formula $\mathrm{E} = \mathrm{kT}$. The new 2019 definition reads:

> The kelvin, symbol K, is the SI unit of thermodynamic temperature; its magnitude is set by fixing the numerical value of the Boltzmann constant to be equal to exactly 1.380649×10^{-23} J/K [joules per kelvin].

As a matter of fact the most accurate measurements of k (about one part in a million) have been obtained by acoustic thermometry, which relies on the fact that the speed of sound in a gas is directly dependent on its temperature.

Can we change? Yes, we can! What has happened over the last half-century is that we have been replacing units defined by certain sacred artefacts kept in highly-esteemed institutions, with units based on precision measurements of certain universal constants or combinations thereof.

The diagram depicted in Figure I.3.2 gives a comprehensive scheme of the newly proposed definitions of the base SI units. The proposal was prepared by the *Comité international des Poids et Mesures* and was officially adopted in 2019. This is quite a substantial upgrade, much like the upgrades of your computer software, except that I would guess that here we talk about version 26 or so, because the first versions go back to about 1875. The base units are represented as colored nodes, and the fundamental constants of nature used to define them correspond to the surrounding brown nodes. The grey arrows indicate how the definitions are hierarchically linked to each other. There are seven fundamental units, and therefore seven constants are needed to fix them. The proposal is interesting in that these seven constants are given exact values when expressed in the base units, and therefore this guarantees a consistent set of definitions if we follow the arrows in the appropriate way. To understand how a unit is defined you look at the arrows coming in to the corresponding node and see where they come from. One arrow comes from an adjacent constant of nature and possible others come from

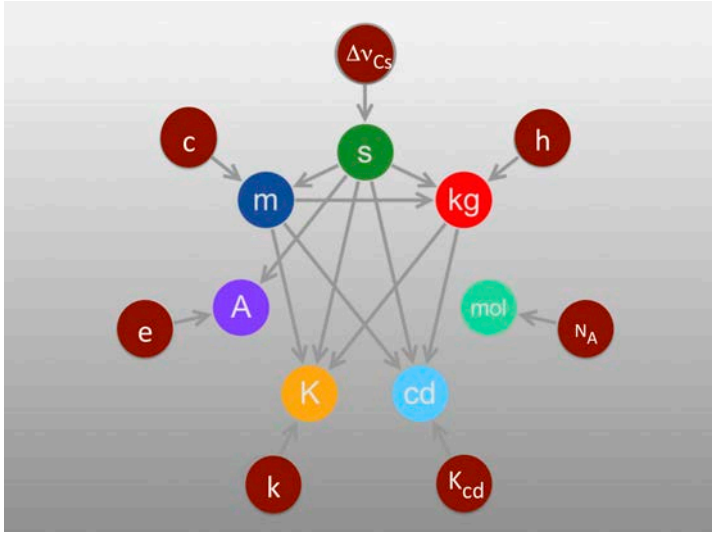


Figure I.3.2: *New SI-Units* The update of the definition of the base SI units adopted by the *Comité international des Poids et Mesures* in 2019. The brown nodes represent integers defining the constants and the arrows indicate the dependencies in the definitions. You start by defining the *second* in terms of the frequency of the ground state Caesium hyperfine transition. Then you move on to the *meter*, the *kilogram* and the *ampère*, all of which involve one additional constant, and then you move on to the *kelvin* and *candela*. (Source: Emilio Pisantly, on Wikipedia.)

units that have been defined before.

Let us consider the definition of the ampère A. We start with the gray arrow coming from the elementary charge e, that arrow represents the exact value of e in terms of A :

$$e \equiv 1.602176634 \times 10^{-19}\ \mathrm{A\ s}\,.$$

The other arrow comes from the 'second,' which is the unit we already defined in terms of the caesium frequency, and therefore A is defined in terms of the observed values of e and ν_{Cs}.

For the kilogram the prototype is no longer used, but reference is now made to the exact value of Planck's constant

h:

$$h \equiv 6.62607015 \times 10^{-34}\ \mathrm{kg\ m^2\ s^{-1}}\,,$$

which now also involves the meter (referring to c and second) and second (referring to ν_{Cs}). So the definition of the kilogram relies on the measurement of the constants h, c and ν_{Cs}.

The remaining question is how, and in what combinations, are these constants determined experimentally. For example magnetic flux Φ that pierces through a two-dimensional superconductor happens to be quantized directly in terms of fundamental constants: $\Phi = n\phi_0 = nh/2e$, from which the *Josephson constant* $K_J = 2e/h$ can be measured extremely precisely. On the other hand, in a so-called quantum Hall system, the Hall-conductivity, which is a transverse conductivity, is quantized in units $\sigma_H = ne^2/h$ that allow for a precise determination of the *Von Klitzing constant*, $R_K = h/e^2$. Measuring these two constants yields an accurate determination of the fundamental constants e and h. Another important observable defined in terms of fundamental constants, which can be measured very precisely, is the *fine structure constant* α,

$$\alpha = \frac{1}{4\pi\varepsilon_0}\frac{e^2}{\hbar c}\,. \qquad \text{(I.3.3)}$$

Indeed the choice of universal constants forms a fair reflection of the depth and precision to which science has managed to descend, and the way they are used in the definition of SI units strikes the optimal balance between precision and reproducibility.

How universal is universal?

Universality is a beautiful, ambitious, but also vulnerable concept, because how do we know whether some constants of nature are universal or not? In mathematics we know such numbers exist, but in physics it is harder to define and establish universality. One should at least require that such would-be universal numbers are the same throughout the (our) universe. But how do we know they are or are not?

Universal constants are – if not God-given – at least Mother-Nature-given-numbers. They happen to be equal to what they have been found to be in human experiments. Their values are believed to be universal, that is, independent of space and time. As you know too well, that doesn't hold for all Mother-Nature-given-numbers, like todays value of your body-mass index for instance, or the viscosity of some expensive French Cognac. If I use phrases like 'the same everywhere and for all time,' I in fact mean everywhere and for all time in our universe, or even better, just nearby in our universe in our present age. Because if we happen to live in a *multiverse* – and there is no fundamental reason why not – then one of the clues about multiverses is that in each separate universe the laws of physics could be quite different. They would represent very different points in the space of possible theories that we have come up with so far. This would imply that there might be entirely different sets of universal constants or known constants could take different values.

Fundamental constants as model parameters. A more pragmatic approach would be to postulate that the universal constants are the numerical input parameters that appear in our theories, such as the masses of elementary particles and the strengths of the fundamental forces. The latter, like Newton's gravitational constant and the electron charge, are also called *coupling constants* because they set the strength of the forces between particles carrying mass and/or charge. The very fact that they appear as input parameters means that they cannot be calculated within that theory; their value can only be determined through experiment. And for all we know these numbers are completely independent.

In mathematics we have universal numbers that are ab-

solute as they can be rigorously defined. The number π for example is defined as the ratio of the circumference of a circle and its diameter. It is a dimensionless number that cannot change and is absolute within the framework of mathematical axioms. One might be tempted to link the dimensionless ratios of physical universal constants to an expression in terms of the universal numbers of mathematics only, much like Plato in his cave would have liked it. In spite of the fact that there is quite an industry actively pursuing these ideas, I consider that somewhat premature. I can only envisage such a step as a final one where the ultimate unified physical theory would be obtained. But nobody promised us such a paradise in the first place so let's go back to the parameters in our current fundamental physical theories.

Reducing the number of fundamental constants. From the perspective of physics it makes complete sense to ask how fundamental these would-be fundamental constants really are. Over time, physical theories get more and more unified in their description of physical phenomena, implying that fewer theories with a smaller number of parameters suffice to account for the same or an even larger body of experimental data. This means that the number of independent fundamental constants has to decrease because we discover relations among them.

Think for example of Maxwell's theory unifying the description of electricity, magnetism and light into a single framework. That theory has in fact three fundamental constants (i) the dielectric constant of the vacuum ϵ_0 featuring in the Coulomb law that gives the force between two electric charges (ii) the magnetic permeability of the vacuum μ_0 featuring in Ampère's law that gives the force between two current carrying wires and (iii) the velocity of light c. Now it turned out that there is a relation between these constants that follows from Maxwell's equations, that relation is just $\mathrm{c} = 1/\sqrt{\epsilon_0\mu_0}$, and it is this relation which allowed us to write the Maxwell equations (I.1.26), with only the velocity of light appearing in them. This is a nice illustration of the fact that the more unified the perspective, the lower the number of independent fundamental constants. This insight forces us to accept that our universal constants are not so universal after all, and it makes us wonder where this game will end.

Where do we stand? Constants that at present are considered to be universal are for example the strength of the gravitational and electric forces $\mathrm{G_N}$, and $e^2/4\pi\epsilon_0$, the velocity of light c, Plank's constant $\hbar$, and Boltzmann's constant k. These constants are *dimensionful*; they are not pure numbers like π, because they have some units linked to them, like c has units *length/time*. That may disappoint you because we are talking about universal constants and they change already if we go from measuring lengths in meters to lengths in inches and the like.

But the good news is that they, exactly because they have units, provide universal – Mother Nature given – links between those different types of units. Such links allow you to eliminate specific units, for example we can use c to convert to units where spatial distance is measured in seconds, $light\ seconds$ to be precise. A distance of one light second is defined as the distance a light pulse would travel in one second, so generally the distance d in $meters$ corresponds to a distance $\mathrm{d/c}$ in $light\ seconds$. This is what we discussed extensively in the previous section. In these units the sun is eight light minutes away while the Andromeda galaxy 2.5 million light years. Planck's constant h appears in the fundamental relation linking energy and frequency postulated by Einstein reading $\mathrm{E} = \mathrm{h}\nu$, and has units $joule \times second$, the velocity of light links mass and energy ($\mathrm{E} = \mathrm{mc}^2$) but also space and time as we saw. Boltzmann's constant links temperature to energy through the relation defining the thermal energy $\mathrm{E} = \frac{1}{2}\mathrm{NkT}$. Having all these relations we could do away with all conversion factors, meaning that you can choose units in which the universal constants (h, c and k) would become equal to unity, and then measure everything in powers of *only* $joules$ (energy) *or* only $meters$ (length) *or* only $seconds$

(time). We will come back to this system of 'natural units' shortly.

Time dependence of fundamental constants? The comments made so far suggest that we take a more pragmatic stand on this question of universality. On a deeper level the value of many would-be universal constants could for example depend on some underlying, hitherto unknown dynamical mechanism, which typically means that they are probably *not* constant in space and time. Instead they are like fruit or peanut butter, in that they have an expiration date. They turn from external input parameters of the old theory into calculable output parameters of the underlying new theory. They move from the pool 'fundamental' to the pool 'effective.' But if this is the way it works it suggests that we should go out and measure whether there are universal constants that do actually vary in time and space. We know for example that the fine structure constant $\alpha = e^2/4\pi\hbar c$ sets the scale for the separation of lines in the atomic spectra, and one could try to make observations of the spectra emitted from atoms that are very, very far away in the universe and check whether the fine structure constant was exactly the same or different at the time the signal was emitted. Experiments of this nature were proposed by John Barrow et al. in 2002. The results of such experiments have so far not confirmed the idea but did produce some upper limit on the relative shift of α of 10^{-17} per year in 2008.

The narrow window of opportunity for life. It is the set of values that these constants of nature have, which turns out to be essential for *our* universe to be what it is. How do we know? Can we go to other universes to check this out? No, not quite, but having reliable theories in which these numbers feature allows us to ask what would have become of our universe if the parameters had had different values. The result of such an exercise is quite surprising not to say startling: it is only in a very narrow window of parameter values that a universe like ours, with its structural complexity and diversity as expressed through the chemistry of life for example, would be possible. We have touched upon some of these aspects in the section about Big Bang cosmology. And others will be mentioned in a section on the ascent of matter in Chapter III.1.

What to do if somebody tells you that they weigh 10^{52} Hertz? If you befriended a music lover and they tell you that their mass is 10^{52} Hertz (1 Hz $=$ 1 inverse second), then you might want to call them crazy, but if they know about universal constants what they say may make complete sense. You can always go back and restore the more familiar units by multiplying with a particular simple combination of fundamental constants. In this case you start with inverse seconds and want to get back to kilograms: $\mathrm{M} = 10^{52}[\mathrm{second}^{-1}] = \mathrm{M} \times \mathrm{h}$ $[\mathrm{joule}] = \mathrm{M} \times \mathrm{h} \times \mathrm{c}^{-2}$ $[\mathrm{kg}]$. So, the upshot is that the combination hc^{-2} converts $[\mathrm{sec}^{-1}]$ into $[\mathrm{kg}]$. The numerical factor involved equals $6\times10^{-34}/9\times10^{16} = 0.66\times10^{-50}$ $[\mathrm{sec\ kg}]$. So having a mass of 10^{52} Hz is actually quite OK. Indeed, units are a matter of convention; if somebody on a market ordered 50 $\mathrm{troy\ ounces}$ of Gouda Cheese, you would not be surprised if I told you that this person was an English jeweler honeymooning in Amsterdam, would you? □

Turning the argument around one could say that choosing the values of the universal constants at random, the chance to end up with an inhabitable universe would be vanishingly small. We expect universes equipped with fancy observers like ourselves to be extremely rare. Lucky us! The *anthropic principle* – a philosophical principle – refers exactly to the attempt to apply the arguments just presented in the opposite order. It tries to derive the structure of our actual universe solely from the fact that we, *homo sapiens*, are here. In a qualitative sense this is of course an interesting question, but as a quantitative ap-

proach it strikes me as naive and doomed. Think of the calculation from quantum first principles, of the anomalous magnetic moment of the electron, which agrees with experiment to twelve significant decimal places! It is hard to imagine getting such precision out of a qualitative approach like the anthropic principle. To understand the universe you need to use far more facts from nature than our mere existence.

Theories outside their comfort zone

Scientific progress can be measured by how effective our theories are. The more physics we explain with the fewer theories, the better. In this section we are going to play some heuristic games with numbers. The observed numerical values for our universal constants tell us what the relevant scales in nature are. At the same time these numbers provide insight in the domains of validity of some of the well-established theories. Surprisingly, naive reasoning and dimensional analysis leads to suggestive qualitative insights with respect to fundamental physics. These arguments underscore the value of heuristics. We have listed some of the fundamental scales with the formulas related to them in Table I.3.2 on page 147.

Domains of validity. Given the values of the universal constants, it is enlightening to cook up other numbers from them which in turn can be interpreted as *characteristic scales* that play a significant role in our universe. Such scales not only follow from the observed values, but also from assumptions underlying the theories in which they appear as parameters. This number cooking game often involves extrapolating the 'laws of nature' to uncomfortable extremes and exactly for that reason this game can yield some information on what the *domain of validity* of such theories really is.

Some devil's advocate, a malign adversary or even a bright student may within the context of a certain model come up with some well-defined, yet, really nasty questions. Questions, which the theory may fail to answer correctly, or may cause the theory to get stuck in a recursive loop that points to a profound confusion or persistent contradiction in our current understanding. Contradictions of a type that faithful teachers sometimes hide, ignore, or even deny. Yet, there always appears to be a moment of truth when it is no longer possibly to deny that the theory fails to give a straight answer to a straightforward question, not even in principle.That is why such Q&A sessions are worth pursuing in spite of their heuristic if not speculative nature. Fortunately many of the theorists I met in my life were always willing and – even eager – to randomly 'shoot the breeze' and ask creative 'what if' questions.

This freedom to let the collective mind wander should be cherished as it is at the heart of scientific progress. And scientific progress is basically about pushing the limit on the ranges of the validity of theories further and further. After each turning point or paradigm shift, the new theory usually provides clear-cut quantitative restrictions on the domain of validity of the old theory; that is why we can speak of scientific progress in the first place[2].

The virtue of heuristics

All we need is the back of an envelope.

Do electrons love or hate each other? We have so far discussed some aspects of the classical theories and some of the salient features of the relativity and quantum domains. And we have commented on the universal constants of nature that we have measured and that feature as

[2] Some devil's advocates therefore argue that particular religions, as systems of knowledge, lack an internal mechanism or stimulus through which they might learn about their limited domain of validity. It is my opinion that the imperative of open questioning and self-improvement sets science apart in the history of human endeavors.

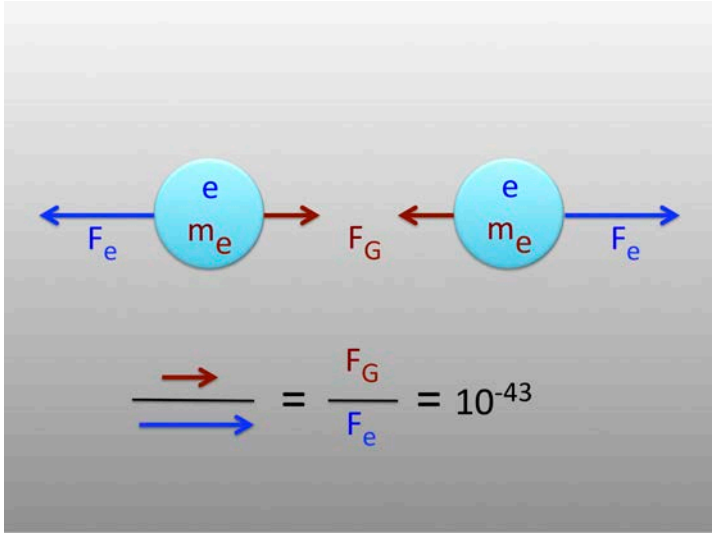


Figure I.3.3: *Interacting electrons.* Two electrons in outer space repel because of their equal charges and attract because of their masses. Yes, and they do fly apart!

external input parameters in our models, like the strength of certain forces and the masses of certain fundamental particles. Combining such numbers and the simple laws in which they appear gives interesting information about the characteristic scales that we observe in nature. That information is at best qualitative and heuristic, but it does provide useful insights about the expected domain of validity of our theories.

In this subsection we limit ourselves to the electromagnetic and gravitational force and what we can conclude from them with respect to the scales that we should associate with them, then we will add some quantum wisdom to it. These two forces are remarkable in that they both have an infinite range and the laws describing them are so-called 'inverse square laws.' For the gravitational force between two masses we have Newton's law, while for the electric force between two charges we have Coulomb's law:

$$F_G = -G_N \frac{m_1 m_2}{r^2} \quad \text{and} \quad F_e = \frac{1}{4\pi\epsilon_0}\frac{q_1 q_2}{r^2}\,. \qquad \text{(I.3.4)}$$

One might ask: Is there a way to compare these forces? Yes and no; they talk about essentially different things like masses and charges, so it's like comparing apples and pears. However, it is not as bad as that because nature has given us particles that have both mass and charge – they are both apples and pears so to speak – and these allow us to compare the strengths of the two forces in a meaningful way. In Figure I.3.3 we show two electrons which have charge e and mass m_e. They attract because of their masses and they repel because they have equal charges. If they met in outer space they would experience two opposite forces, so the key question is: will they pair up or fly apart? To get the answer, we have to take the ratio of the magnitudes of the two forces,

$$\frac{F_G}{F_e} = \frac{4\pi\epsilon_0 G_N m_1 m_2}{q_1 q_2} \simeq 10^{-43}\,. \qquad \text{(I.3.5)}$$

This shows that the gravitational force is phenomenally weaker than the electric force. Note that this ratio does not depend on the distance; it is a fixed number. How sad for the electrons, it is not only hard to stay together; it would even be extremely hard to meet in the first place. The order of magnitude of this number holds for any fundamental particle which carries both mass and charge, though the actual number could differ of course.

An electromagnetic size: how big is an electron? Given the force between two charges one can calculate the interaction energy of two charges that are separated by a certain distance. One may also define what is called the *seif-interaction energy* of a particle due to the force field. This electrostatic self energy is the energy it costs to build up a charge e on a sphere of radius r, and is of the order of $e^2/4\pi\epsilon_0 r$. Building up a charge means that you bring in infinitesimal amounts of charge from infinity and calculate the interaction potential. Equating that potential energy to its mass energy $m_e c^2$ according to the famous Einstein formula yields the *classical electron radius* in terms of its mass:

$$r_e = \frac{e^2}{4\pi\epsilon_0 m_e c^2} = 2,8 \times 10^{-15}\ \text{m}.$$

This expression is directly obtained from combining certain constants of nature known from experiments with naive dimensional analysis, and begs for an interpretation even though it would be heuristic. It certainly is a size one can naturally assign to a charged particle as it reflects the energy of the total electric field carried by a charge on a sphere of radius r_e. Note that the electromagnetic size of a charge grows with charge but decreases with increasing mass. So the lightest particle having a certain charge yields an upper-bound for the electromagnetic size of such a charge. Note the paradoxical nature of this classical reasoning, it would produce an infinite potential if one would assume the particles to be point like. This fact stood out as a fundamental limitation of the classical theories in the description of a charged particle, and as we'll see in Volume III, Chapter III.4, quantum field theory provided an essential new perspective on this question that exploited the sophisticated notion of *renormalization*. Anyway, according to the reasoning we have followed so far, a particle of zero charge could still be considered point-like.

A gravitational size: know your horizons! As for a neutral particle electromagnetic considerations are void, so one could maybe make use of the gravitational interaction to set a scale, and assign a classical gravitational radius to any mass m. One repeats the argument and replaes Coulomb's law by Newton's gravitational law, ignoring for the moment the sign difference[3], so the potential energy of a mass m at radius r would be $E \sim G_N m^2/r$, and equating this to the mass energy $E = mc^2$, we get $r_g \sim G_N m/c^2$. This relation sets a scale for the applicability of classical Newtonian gravity, and indeed, remarkable enough it is (up to a factor 2) equal to the *Schwarzschild radius* of a particle of mass m defined as:

$$R_s = \frac{2G_N m}{c^2}\,; \tag{I.3.6}$$

that is $\sim 10^{-57}$ m for the electron. This is an excruciatingly small number, far outside of the scope where our physical intuition has any experience, let alone any bearing. It's like somebody getting up and starting to talk to you about what they are planning to get done in the next one billionth of a second! Stay normal please! The Schwarzschild radius is where the gravitational horizon around a black hole with mass m is located, and according to the general theory of relativity, there is no information that we, as outside observers, can obtain about the interior of the black hole. Talking about a particle's properties beyond that scale is problematic. If you would send a willing observer to check out the interior they would not be able to report back to you, as they are doomed to a not so gracious exit facing the singularity at the origin.

No escape: apocalypse you! To clarify this peculiar property of black holes, it suffices to repeat the thought experiment that the French mathematician Pière-Simon Marquis de Laplace described in 1796, and that lead him to the notion of the *corps obscur*, which in modern parlance is just a black hole[4]. You probably are familiar with the notion of *escape velocity*, if you throw this book straight up in the air it will under most circumstances drop on your head some time later. Yet, if you throw it with a speed of more than 11 kilometers per second, then it would never return. As you see it is not so simple to get rid of a book, they tend to stick around. Far away it would still feel the gravitational force caused by the mass M of the Earth, but it can escape because the kinetic energy would be larger than its gravitational binding energy to the earth. Equating the kinetic energy and the binding energy gives the equation for v_{esc}, we obtain:

$$m v_{esc}^2/2 = m G_N M/r \;\;\Rightarrow\;\; v_{esc} = \sqrt{2G_N M/r}\,. \tag{I.3.7}$$

Note that this velocity does not depend on the mass m of the book, so anything you throw up with a velocity exceeding 11 km/s will be gone for ever. You see that the

[3] The sign would translate in the statement that the the potential corresponds to the energy needed to gradually bring the mass to infinite radius.

[4] A British natural scientist, John Mitchel, had already made a similar argument in 1783. He called the objects *dark stars*.

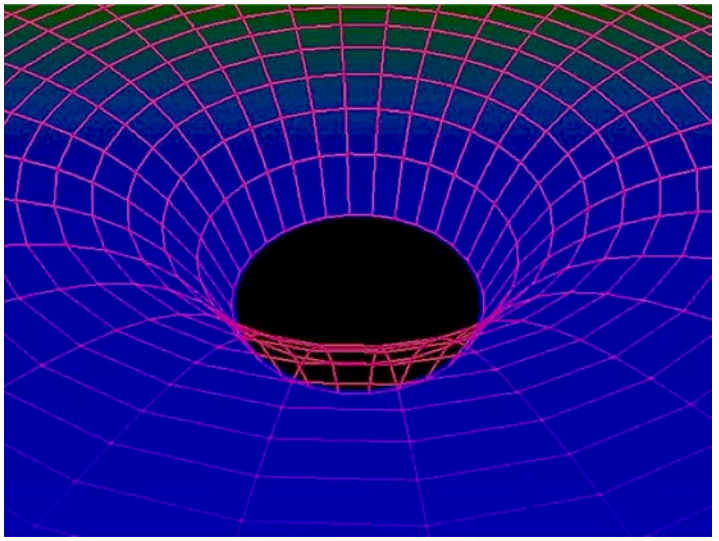


Figure I.3.4: *A black hole.* General relativity tells us that if we put a lot of mass in a tiny volume that mass will collapse under its own weight and form a black hole. Black, because its escape velocity exceeds the speed of light and – at least classically – no information can escape. A virtual sphere called the *event horizon* will form outside of the mass, and its radius corresponds to the Schwarzschild radius (I.3.6).

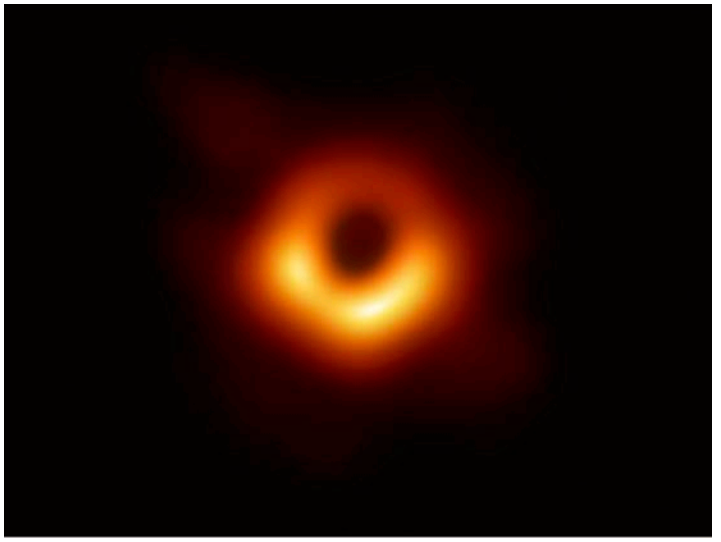


Figure I.3.5: *A black hole picture.* This is a real picture of a black hole in the galaxy M87 about $5 \times 10^{20}\mathrm{km}$ away. It measures $4 \times 10^{10}\mathrm{km}$ across, and has a mass corresponding to 6.5 billion solar masses. The picture was captured by the Event Horizon Telescope (EHT), a network of eight linked telescopes on Earth.

escape velocity would increase if we would decrease the radius of the Earth while keeping its mass fixed. And indeed, knowing that the velocity of light was approximately $300.000\ \mathrm{km}/s$, Laplace basically asked himself the question: to what radius do we have to shrink the size of the Earth in order that the escape velocity would become equal to the velocity of light? And beyond that radius, he argued, even light would not be able to escape from the Earth's surface – the Earth having the size of a marble by the way. No light signals could be sent to some far away observer, at least they would not get very far. The Earth would be black: a black hole so to speak. Though this tiny Earth would be invisible, you would still be able to probe its presence gravitationally. If the Sun were a black hole, you wouldn't be able to see it but the planets would move in their orbits all the same. Going back to the formula (I.3.7), you'll also agree that an object with *any* given mass M will have an event horizon once its size is small enough. With black holes one tends to think of super massive objects like heavy stars. After having burned up their nuclear fuel they would collapse under their own gravitational attraction in a supernova event. The compact object staying behind would indeed be a black hole. Astrophysicists have in the meantime identified large numbers of them. They also are located at the center of galaxies. There one suspects the presence of a giant black hole gobbling loads of stars for breakfast. At first nobody could think of compactifying a chunk of matter like the Earth to within a radius smaller then its horizon of less than a centimeter. The concept of a black hole was so totally inconceivable that it was discarded as a brilliant fiction of the mind – clearly an artefact of fancy mathematics. The idea was that a condensed state of matter, like the space inside of a stone or a lead block, would be 'filled up' completely. It would be only compressible to a limited extent, which seemed evi-

dent from just experimenting with it. There would be no room for such extreme collapses was the prevailing opinion, which was even held by influential astrophysicists like Sir Arthur Eddington.

In fact the story went the other way around. With the advent of the quantum understanding of the deep structure of matter, it was that intuitive idea that matter fills space, which turned out to be a fiction of the mind. Quantum theory taught us exactly the opposite, that matter *is* mostly empty space. The mass of a stone is carried mostly by the tiny nuclei inside the atoms and the spacing between those nuclei is about a million times larger than their size. Removing that space, you could in principle compress the Earth to the size of a meter across – the density would then correspond to the density of a neutron star. Astrophysicists have systematically studied the processes of stellar evolution, including their dramatic ending. A star will, depending on its mass, end up as a compact object, like a white dwarf, a neutron star, or a black hole. Most black holes observed have masses between five and several tens of solar masses, and the lightest known black hole has a mass of around 3 solar masses.

A second comment to make is that small masses also have a horizon, which makes it possible to study mini black holes in order to find out to what extent they could be produced and would be stable. Maybe also these hypotheticals – fictions of the mind – will be found one day in spite of them being 'invisible.'

The age of our universe. Before continuing our black hole adventure where science is running into at least one of its own horizons – if not a brick wall – we briefly return to the other side of General Relativity (GR) connected with the Friedmann cosmology, which we discussed quite extensively in the previous chapter. The question we want to address is a question that has puzzled humankind already for millennia, but at the same time it is also a question that children start asking when they are in elementary school. Did the world always exist, or was there a beginning – a moment of creation? And if so, when was that? Such questions that everybody encounters at some point in their life create a demand for answers, and where there is demand, economists tell us that there will be supply. And so there was!

There is a great history of estimating the age of the universe. The early estimates from a smart clergyman who managed to argue from The Scriptures that the week of creation was about 4000 years ago are well known. The story is that the Bishop James Ussher around 1650 came even with a precise date: Sunday 23 October 4004 BC! What you can say about the history of would-be answers is that there was an overall trend to ever increasing numbers.

It is interesting to recall the involvement of the great biologist Charles Darwin who estimated the age of the earth by using geological arguments combined with the time needed to have the complexity of life evolve, to be a few hundred million years. This estimate was heavily criticized by Lord Kelvin who argued that the age of the sun, based on the state of knowledge – or ignorance – of the day, could not be more than say 20 or 30 million years. His knowledge typically comprised Newtonian gravity, chemistry and thermodynamics, and his ignorance was hidden in the fact that he didn't know that he didn't know. The unknown unknowns concerned the whole field of nuclear physics, because there was none in those days. And to understand the age of the Sun you have to understand the nuclear processes that keep the Sun shining. This was an exemplary scientific debate, where Darwin got much closer to the correct answer for reasons that are clear now. In the second half of the twentieth century the astronomers entered the game using a variety of observational and calculational methods. This caused the numbers to go up dramatically into billions of years. Fortunately the results also started to converge.

Let us try to make a crude estimate of the age of the universe starting from the Friedman equation (I.2.9). For simplicity we assume that the universe is flat $(k = 0)$ and has only matter in it. The matter density drops inversely with the volume, so:

$$\rho_m = \rho_{crit}\frac{\Omega_m}{a^3}.$$

with ρ_{crit} the critical energy density and Ω_m the present relative matter constant. The equation then simplifies considerably and we get:

$$\sqrt{a}\frac{da}{dt} = H_0\sqrt{\Omega_m}\,. \tag{I.3.8}$$

As you can check by differentiating, the solution is $a = \beta t^{2/3}$ where β is some constant. This yields for the Hubble parameter $H(t) = 2/(3t)$. Evaluating this for t_0 we obtain $H_0 = 2/(3t_0)$, resulting in the estimate,

$$t_0 = \frac{2}{3H_0} \simeq 9.3 \times 10^9\ \mathrm{yr}\,,$$

where we have used the value $H_0 = 70 \mathrm{km s^{-1}\ Mpc^{-1}}$. This crude calculation thus shows that the age of the universe is of the order of the inverse Hubble parameter. The best value available today, extracted from the 2018 data of the Planck space telescope is:

$$t_0 = (13.781 \pm 0.020)\ \times 10^9 \mathrm{yr}\,.$$

Note the amazing precision here, which shows the tremendous progress in the field of observational cosmology! This means that the Hubble parameter is a fundamental observable as it sets the scale for the age of the expanding universe.

Going quantum

The quantum size of a particle. So far we used the equations of classical physics and relativity, which involved the fundamental constants G_n, e, and c. What happens if we include some of the basic quantum relations? This would add Planck's constant h (or its reduced version $\hbar = h/2\pi$, denoted as 'h-bar') into our deliberations.

A nice starting point is the expression that Louis de Broglie[5] in 1923 proposed for the wavelength λ of the 'matter wave' associated with a particle of mass m moving with velocity v or momentum $p = mv$, which simply reads $\lambda = h/mv$. Combining this formula with Einstein's dictum that nothing can move faster than light implying that $v \leq c$, we arrive at a 'minimal wavelength'

$$\lambda_c = \frac{h}{mc}\,, \tag{I.3.9}$$

for a quantum particle, which is called its *Compton wavelength*. The Compton wavelength for the electron is $2.43 \times 10^{-12}\ m$, which on a heuristic level can be interpreted as a measure for the 'quantum size' of the electron. For scales much larger than the Compton wavelength we can safely consider the electron as a well-defined localized 'particle' whereas when we approach the Compton wavelength we have to take its wavy nature into account and treat it quantum mechanically. In other words also in quantum theory the notion of a point particle breaks down beyond a certain scale. A rigorous way to define the Compton wavelength is to say that it equals – following Einstein – the wavelength of a photon whose energy equals that of the rest energy of a particle: $E = hc/\lambda = mc^2$. This is certainly true but less straightforward to interpret.

Alternatively we may invoke Heisenberg's uncertainty relation $\Delta x\ \Delta p \geq \hbar/2$, which in words amounts to the statement that in a given quantum state of a particle the uncertainty in the outcome of a position measurement times the uncertainty in the outcome of a momentum measurement equals at least h-bar over two. And if we then interpret mc as the maximum uncertainty in momentum that leads to the Compton wavelength as the minimal uncertainty in

[5] In fact I should have said: Louis-Victor-Pierre-Raymond, the 7th Duke of Broglie!

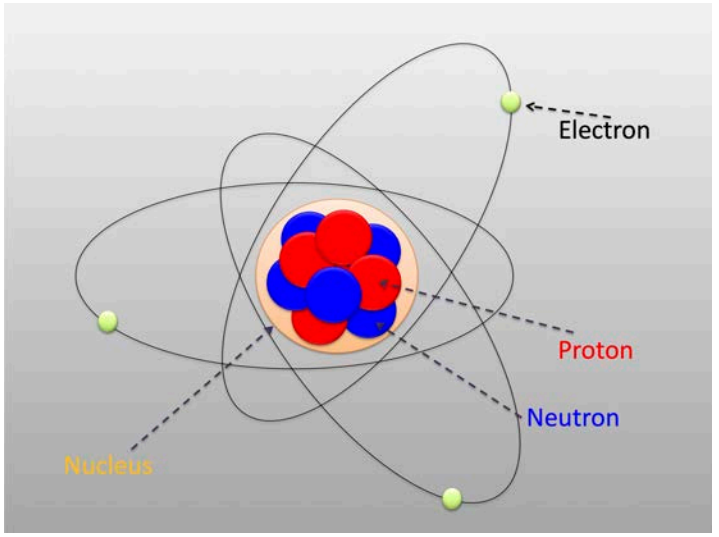


Figure I.3.6: *The Rutherford model of the atom.* The canonical picture of the atom proposed by Ernest Rutherford, with a positively charged nucleus consisting of protons and neutrons, with some negatively charged electrons orbiting the nucleus. It is a symbolic representation, which is misleading in two ways. The relative sizes are totally out of proportion, since the size of the orbits is about 100,000 times larger than the size of the nucleus. So if you take the nucleus as depicted in this figure the electron orbits would be about a kilometer in size! Furthermore, in the stationary states of the atom, the electrons are not at all localized like point particles. The states rather correspond to the 'standing' wave patterns proposed by Bohr as indicated in the next figure. They represent the smeared out probability distributions for finding the electron at a given location.

position of the particle. As we will see later this scale is directly linked to the width of the 'wave packet' representing the electron in quantum theory. And with a quantum leap in vagueness you could argue that minimal uncertainty in position indicates the effective size of the quantum equivalent of a particle.

Heuristics and reasoning by analogy is a dangerous game but can be enlightening and yields a rough sense of the scales involved with little work, not more than that. Therefore very useful!

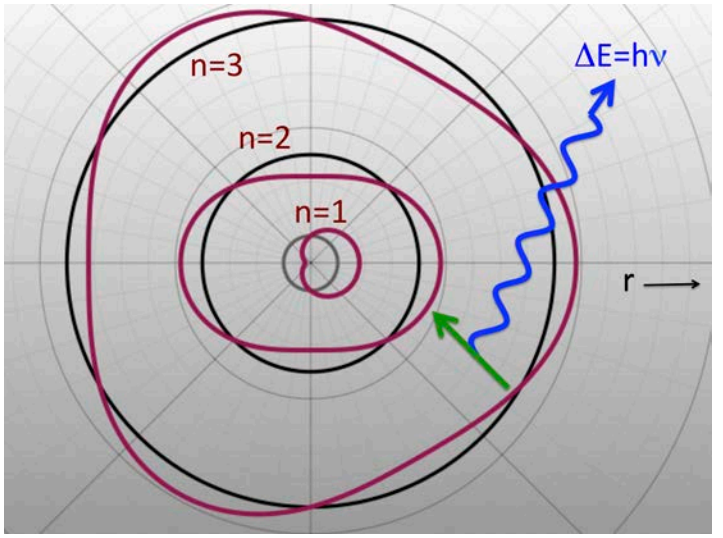


Figure I.3.7: *The Bohr atom.* This model has *quantized orbits*, satisfying the constraint that the electron wave would fit an integer number of times on the orbit. This condition $n\lambda = 2\pi r_n$ leads to states with quantized energy and angular momentum. The radii of the successive orbits scale quadratically ($\sim n^2$).

.

An atomic size: the Bohr radius. There is one more quantum scale that we should mention at this point. It is the first quantum estimate of the atomic size called the *Bohr radius*. In 1911 Rutherford had shown that the atom has an almost point-like positively charged nucleus with the electrons orbiting around it. This brought Niels Bohr to his famous atomic model that, with its simple but radical starting point, immediately led to an astonishingly deep insight in the line structure of atomic spectra. Indeed, it is one of the most outstanding results of early quantum theory. The argument used the wave character of the electron (say De Broglie's formula) to quantize atomic orbits, and thereby also its energy levels. From these energies the frequencies of the lines in the spectra could be calculated directly.

Bohr used the idea of *particle-wave duality*, and put it into practice by assuming that the stationary electron states in

the tiny atomic environment would correspond to standing waves on a supposedly classical orbit. That wave should not destructively interfere with itself and therefore Bohr demanded that the electron wave would fit an integer number of times on the orbit of the electron, which led him to the 'quantization condition': $n\lambda = 2\pi r$ with $n = 1, 2, 3, \ldots$. If you now use the relation of De Broglie between momentum and wavelength you get $p = h/\lambda = nh/(2\pi r) = n\hbar/r$. A straightforward exercise in Newtonian mechanics shows that in order to have a circular orbit you need a central force $F_c = ma = mv^2/r$, which in this case is provided by the Coulomb force F_e of equation (I.3.4). So, from the equation $F_c = F_e$ one finds the possible radii r_n :

$$\frac{mv^2}{r} = \frac{p^2}{mr} = \frac{n^2\hbar^2}{mr^3} = \frac{e^2}{4\pi\varepsilon_0 r^2}\,,$$

from which Bohr derived the quantization rule:

$$r_n = a_0 n^2;\ \text{with}\ a_o \equiv r_1 = \frac{4\pi\varepsilon_0\hbar^2}{me^2} \sim 5.3\times 10^{-11}\ m\,,$$

where the constant a_0 in honor of its creator is called the *Bohr radius*. In Figure I.3.7 we have sketched the periodic electron waves for the first few orbits.

It is no surprise that the quantization of the orbits implies that other physical quantities are also quantized, notably the energies and the angular momentum. To start with the latter, for a circular motion we have that the angular momentum $L = rp$, and just substituting the quantized value for p given above, one gets $L = n\hbar$, showing the basic integer quantization condition for orbital angular momentum, which indeed has the dimensions $[kg\ m^2/s]$ of angular momentum. Substituting the radius in the expression for the total energy $E = E_{kin} + E_{pot} = p^2/2m + V_{Coul}$, one finds that the energy is quantized as

$$E_n = E_1/n^2\,, \tag{I.3.10}$$

where the ground state energy is given by:

$$E_1 = \frac{me^4}{32\pi^2\varepsilon_0^2\hbar^2} \simeq -13.6\ eV\,. \tag{I.3.11}$$

We see that the energies of the hydrogen atom are negative (meaning that they are bound states) and that for large n the states pile up towards $E = 0$. An essential feature of the model which, also depicted in Figure I.3.7, is the proposition that when an electron makes a transition from a higher to a lower orbit, the energy difference ΔE will be carried away by a photon that has a frequency $h\nu = \Delta E$. We return to the Bohr model and its relationship with the observed atomic line spectra in the section on atomic structure in the next chapter.

Further gaming with fundamental scales. Returning to typical length scales related to the electron, we have so far cooked up three sizes: (i) the classical electromagnetic size (= the classical electron radius) $r_e \sim 10^{-15}\ m$, (ii) the gravitational radius (= the Schwarschild radius) $R_s \sim 10^{-57}\ m$, and (iii) the quantum scale (= its Compton wavelength) $\lambda_c \sim 10^{-12}\ m$. One thing these numbers clearly suggest is that in worrying about the size of the electron we should first take into account quantum effects, before entering into profound debates on the meaning of its classical electromagnetic or gravitational radii.

What else can we do with these length scales? We could take their ratios and try to interpret them. We can for example define the dimensionless ratio of the (i) and (iii). This number (up to the factor $h/\hbar = 2\pi$) is denoted as α and called the *fine structure constant* $\alpha = e^2/4\pi\epsilon_0\hbar c$; it is indeed a pure number and equals $\alpha \simeq 1/137$. This constant is a clean measure of the interaction strength of the electromagnetic interaction in (relativistic) quantum theory, which is not so surprising because the fundamental constants e, c and $\hbar$ feature in it.

Another dimensionless ratio one can take is the Compton wavelength over the Bohr radius, giving an idea as to what extent the electron would fit in the atom. One finds that $\lambda_c/a_0 = 2\pi\alpha$, which is again proportional to the fine structure constant. This indicates that the two scales are not vastly different, particularly if one takes into account that

the electron in the Bohr-atom is non-relativistic and that the Compton wavelength is an underestimate for its quantum size. This underscores once more that we should treat the problem of atomic structure with quantum theory.

We could also define a gravitational fine structure constant as $\alpha_g = Gm_e^2/\hbar c$, which would equal $\alpha_g = 1.75 \times 10^{-45}$. The ratio of these two 'structure constants', which also equals the ratio of (iii) and (i), brings us back to the intrinsic difference in coupling strength that we mentioned before: the gravitational attraction of two electrons is weaker than their electromagnetic repulsion by some 43 orders of magnitude.

A quantum bound on processing speed. The uncertainty relations allow one to construct heuristic quantum bounds on various dual observables like momentum and spatial extent, or energy and time. The former yielded the Compton length, and the latter allows us for example to set an ultimate bound on processing speed. If we take the energy corresponding to a mass $E = mc^2$ and relate that energy to a fundamental frequency according to $E = h\nu$ and interpret this frequency as the number of logical operations per second, we arrive at the formula proposed by Seth Lloyd in a 2000 paper for the maximal number of transitions N^* per unit mass per unit time.

$$N^* \simeq \nu = c^2/h\,. \tag{I.3.12}$$

Putting in the numbers one arrives at the ultimate processing speed of a 'one kilogram laptop' as some 10^{50} logical operations per second. To give you an idea of what this means: typical estimates for the human brain yield 10^{15}, while the most powerful super computers run at $10 - 100 \times 10^{15}$ flops. These comparisons are rather misleading because of the very different structure of these 'machines;' the brain has a relatively low clock speed of about $100\mathrm{Hz}$ but works in a highly parallel mode.

Nuclear forces: the story of weak and strong. Later on in the book we will discuss two other forces which are not of the inverse square type: the strong and weak nuclear forces They differ in an essential way in that they effectively only act over small distances – meaning, small compared to the size of an atom – and that is why we don't see or feel them. These forces can be approximated by a inverse square law, which is cut-off at a certain characteristic scale, called the strong and weak scales respectively. The effective potential of a weak or strong charge corresponds to the so-called *Yukawa potential*:

$$V_Y = g_Y \frac{1}{r} e^{-r/\lambda_c}\,. \tag{I.3.13}$$

We see that the inverse $1/r$ potential standard for gravity and electromagnetism with a strength g_Y is multiplied with a negative exponential of the distance. The interaction potential is said to be *screened* and becomes vanishingly small past the typical scale λ_c, appearing as an additional fundamental parameter in the theory. Such screening effects make the interactions effectively *short range*. The particle's experience would be comparable to driving in a dense mist or calling each other in the crowd, the interaction between entities is only effective at short distances. We have depicted the Yukawa potential in Figure I.3.8. The interpretation of the characteristic scale λ_c is, that it is inversely proportional to the mass m_c of the particle that is mediating the nuclear force, like the photon mediates the electromagnetic force. The natural relation between a mass and a characteristic length is the quantum scale or Compton wavelength of the particle as pointed out before.

To complete this short interlude on the nuclear forces, let us just give you the scales involved. The strong nuclear scale is in the first approximation associated with the exchange of so-called *pion* particles, their masses are of order $m_\pi \simeq \frac{1}{10} m_p$, yielding a length scale of approximately $10^{-14}\mathrm{cm}$, which typically is the size of a nucleus. For the weak nuclear force the mediating particles are the W and Z bosons with masses $M_W \simeq 100 m_p$ and consequently the weak force has a tiny range of the order of

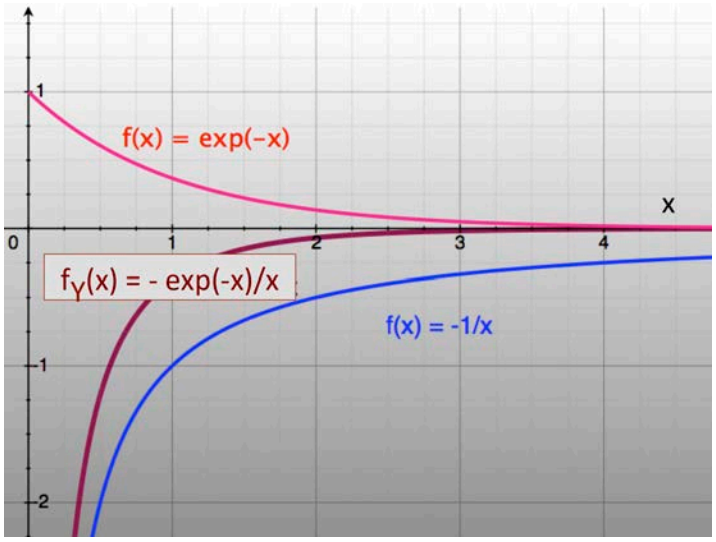


Figure I.3.8: *The Yukawa potential.* The purple curve is the product of a $(-1/x)$ potential in blue (like the one of electromagnetism or gravity) and an exponential suppression factor $(\exp\ -x)$ in red. What results is an attractive potential for the weak and strong interactions, which is effectively short-range because of the exponential cut-off.

$$\lambda_W \simeq 10^{-17}\mathrm{cm}\,.$$

At this point we may also mention that setting the mass of the mediating particle to zero we get back to the formulas of electromagnetism and gravity (the blue curve in Figure I.3.8). This confirms our earlier claim that these long range force fields are associated with the exchange of massless particles such as the photon and the graviton. In that case the potential is a simple power law reflecting the scale free nature of the long range interaction. The power law potential exhibits therefore what in modern parlance is called a *long* or *fat tail*, which refers to the behavior clearly visible on the right in Figure I.3.8 where the blue power law curve is much larger than the exponentially suppressed purple and red curves.

Where the quantum collective rears its head. We started in Chapter I.1 by summarizing the fundamental theories of classical physics, and we have indicated in this chapter how quantum theory enters to indicate the boundaries of the domain of validity of the classical theories of mechanics, gravity and electromagnetism. It will not surprise you that the theory of statistical or thermal physics has also an intrinsic parameter that tells you when quantum phenomena should be expected to become relevant in multi-particle systems such as gases and liquids. The interesting thing here is that these phenomena even occur in 'ideal' systems where we ignore inter particle interactions. The central observation is again based on a simple dimensional argument. If one considers an ideal gas of massive atoms in equilibrium at some temperature T, then the average *thermal energy* per particle is $\mathrm{E_{th}} = 3\mathrm{kT}/2$. So we can define a *thermal momentum* $\mathrm{p_{th}}$ through the relation:

$$\frac{\mathrm{p_{th}^2}}{2\mathrm{m}} = \frac{3\mathrm{kT}}{2}\,. \tag{I.3.14}$$

Next we use the De Broglie relation to define a *thermal wavelength* as $\lambda_{\mathrm{th}} = \mathrm{h}/\mathrm{p_{th}} = \mathrm{h}/\sqrt{3\mathrm{mkT}}$. This length scale depends on h and defines the size of the wave packets related to the thermal excitations of the particles in a gas. For the case of particles in a gas at room temperature the thermal wavelength is typically of the order 0.1 *ängström* or 10^{-11} meters. When does this scale become relevant? It clearly matters if it becomes of the order or larger than the typical inter-particle distance d, which is determined by the particle number density n defined as $\mathrm{n} = \mathrm{N}/V$. Classical considerations (even in the absence of interactions) should break down if:

$$\lambda_{\mathrm{th}} \geq \mathrm{d} \;\Rightarrow\; \frac{\mathrm{hn}^{1/3}}{\sqrt{3\mathrm{mkT}}} \geq 1\,. \tag{I.3.15}$$

The conclusion is that we enter the quantum domain at high density and/or low temperature. As we will discuss later on, this is intimately linked to spectacular quantum phenomena like superfluidity, (super-)conductivity, and Bose-Einstein condensation. A rough indication of some examples where the quantum laws are inescapable can be found in table I.3.1.

Table I.3.1: Thermal wavelengths and domains*

System	T[K]	λ_{th}/d	domain
Air at room temperature	300	0.006	classical
Liquid nitrogen (^4He)	77	0.10	classical
Liquid helium (^4He)	4	1.16	quantum
Electrons in copper	300	18.9	quantum

*R. Baierlein, Thermal Physics, Cambridge Un. Press (1999).

Natural units ©1898 Max Planck

We conclude by discussing a system of natural units introduced by Max Planck. The price to pay is to give up anthropocentricity, at least on the level of units.

Can we please finish this endless talk about units? Yes, we can! It was already pointed out by Max Planck how to do this in the *Fünfte Mitteilung*, *Über irreversible Strahlungsvorgänge*, to the *Preussische Akademie* in 1898[6]:

> Alle bisher in Gebrauch genommen physikalischen Maßsysteme, auch der sogenannte absolute C.G.S.-System, verdanken ihren Uhrsprung insofern dem Zusammentreffen zufälliger Umstände, als die Wahl der jedem System zu Grunde liegenden Einheiten nicht nach allgemeinen, notwendig für alle Orte und Zeiten bedeutungsvollen Gesichtspunkten, sondern wesentlich mit Rücksicht auf die speziellen Bedürfnisse unserer irdischen Kultur getroffen ist.

In the *Mitteilung* he devised a system of units that deserves the qualification *natural* like no other. These *Planck-units* are all directly linked to the simple universal constants that we discussed before:

– the gravitational constant G_N with units $[kg^{-1}m^3s^{-2}]$,
– the speed of light $c \sim [m\,s^{-1}]$,
– Planck's constant[7] $\hbar \sim [kg\,m^2s^{-1}]$
– and Boltzmann's constant $k \sim [kg\,m^2s^{-2}]$.

Some juggling with dimensions leads quite unambiguously to the following natural units: the Planck-unit of length or the Planck length,

$$l_p = \sqrt{\frac{\hbar G_N}{c^3}} = 1.62 \times 10^{-33}\ cm;$$

the Planck mass,

$$m_p = \sqrt{\frac{\hbar c}{G_N}} = 2.18 \times 10^{-5}\ g;$$

and the Planck time,

$$t_p = \sqrt{\frac{\hbar G_N}{c^5}} = 5.39 \times 10^{-44}\ s.$$

If we include the Boltzmann constant k as another fundamental constant, we may add the Planck unit of temperature:

$$T_p = \sqrt{\frac{\hbar c^5}{k^2 G_N}} = 1.42 \times 10^{32}\ K.$$

Divine units indeed! Imagine, adopting these as the units of length, mass, time and temperature amounts to setting all the above expressions in terms of the fundamental constants equal to one, which implies that we have to set $\hbar = c = k = 1$ in all formulas and calculations! What a relief for the students who have to remember them. I am afraid though that in the real world of construction and electrical engineers these units would be despised

[6] English translation (by author): All physical systems of measurement, including the so-called absolute CGS system, which have hitherto been used, owe their existence to accidental circumstances, in that the choice of the units on which each system is based does not depend on general points of view that necessarily hold for all places and times but takes only in consideration the special needs of our earthly culture.

[7] In Planck's original paper this or better *his* constant was actually called b and not $\hbar$.

except in the rare instance where one is involved in building universes[8]. This is precisely the case because using these divine units would involve such huge conversion factors that you would lose a common sense of scale.. On the other hand, dragging along all these fundamental constants all the time makes formulas far less transparent and that clutters the mind. I challenge the entrepreneurial readers to choose natural units for the rest of this chapter, which means that you set the universal constants everywhere equal to one. You will find that the resulting formulas become stunningly simple indeed.

Ahead of the crowd. The natural units beg for an interpretation and maybe it is just that they mark the domain of validity of the theories of Einstein and/or quantum theory. Or better, they mark a domain where quantum and relativity typically meet. Problems at the Planck scale involve phenomena where quantum gravitational effects have to be included. And if we do not have a complete understanding of what the quantum theory of gravity is, our calculations will be unreliable to say the least, and may give unsatisfactory answers to sensible questions. Referring back to the scales we discussed before we see that for a fundamental particle with a mass equal to m_p, the Compton wavelength and the Schwarzschild radius become roughly equal since:

$$\hbar/m_p c = G_N m_p/c^2.$$

This expresses the fact that for a particle with a mass of the order of the Planck mass the quantum uncertainty in its spatial extent is the same as its 'gravitational' uncertainty. This gravitational uncertainty is due to the strong gravitational field which causes that it is impossible to extract information on the outside of that tiny horizon about what happens inside. The equal sign in the above equation, inspired by matching uncertainties, basically makes the bold hypothesis that both uncertainties are somehow due to the same underlying mechanism. Such a mechanism would have to be accounted for by a would-be theory of quantum gravity.

It is worth remembering that heavy objects have a Compton wavelength that is negligible, for example for the earth we get that $\lambda_\oplus = \hbar/m_\oplus c \simeq 10^{-67}\ cm$, while its Schwarzschild radius still is a respectable $R_\oplus = 0.9\ cm$. And because both are so much smaller than the actual size of the object Earth, it is not in terrestrial physics that this fundamental contradiction will leave any mark. For the electron the situation is the opposite, $\lambda_e = \hbar/m_e c \simeq 10^{-12}\ cm$ (indeed its non-local character manifests itself on the atomic scale), while the Schwarzschild radius is an excruciatingly small $R_e = 2G_N m_e/c^2 \simeq 10^{-57} m$. The conceptual conflict between relativity and quantum theory as encountered at the Planck scale signals the crisis our notion of space-time suffers in the light of the quantum postulates. On the other hand, one might hope that also this crisis will be the seed for a new fundamental paradigm.

Black holes

> The question is not whether black holes 'exist'. They exist as classical objects. The question is, what is the quantum mechanical equivalent? It is well possible that in quantum mechanics black holes are no longer strictly distinguishable from more conventional forms of matter.
>
> Gerard 't Hooft, *Physica Lecture* (1995)

It is widely believed that black holes are rather esoteric, farfetched, out of this world, nerdy gadgets and therefore not so relevant. Wrong! It has become ever more clear that they are the principal key to a new and much deeper understanding of what gravity and thus space-time are really about. It introduced the concept of *information* into physics in a fundamental way. Indeed, some people say that

[8] Surprisingly, quite a few engineers appear to do so in their spare time. I would rather have engineers constructing universes, than philosophers building airplanes!

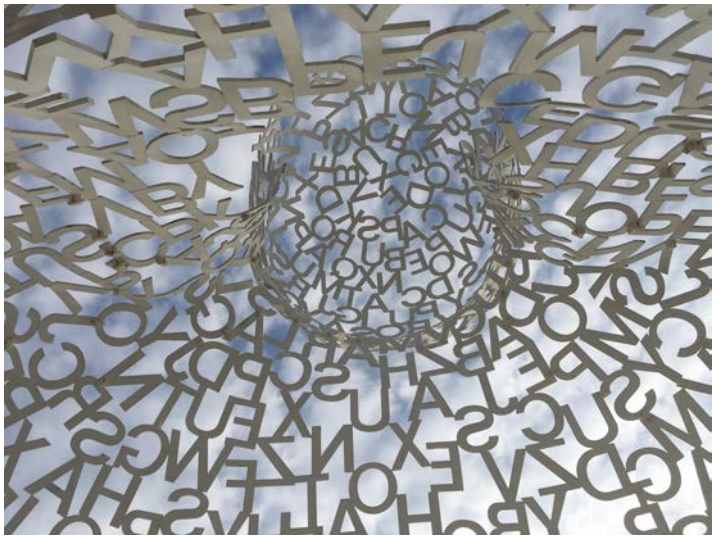


Figure I.3.9: *Information loss?* Does information get lost forever when it falls into a black hole? Detail of the remarkable sculpture *Le Nomade* in Antibes (France) created by the Spanish sculptor Jaume Plensa

'black holes' are for the theory of gravity, what the 'hydrogen atom' was for quantum theory. Gravity's fundamental properties and problems really show up in the unexpected intricacies of black hole geometry. There is more to gravity than dropping a teaspoon on the floor, or keeping the moon in orbit. We have already seen in the previous section the peculiar property of horizons predicted by GR. But just saying that there is a horizon is not sufficient. When you start thinking about it seriously, a lot of hard questions come your way, questions that probe the deeper grounds of GR and go beyond it. This section touches upon the remarkable research by many of the brightest brains of recent generations, attempting to bridge the gap between curved space-time and quantum theory. This appears necessary to get a complete and consistent picture of these miraculous outposts of reality. Yes, horizons mark an important frontier, but to what?

Stephen Hawking: Quantum black holes are not black! We have discussed the essential feature of GR that every mass M has a Schwarzschild radius $\mathrm{R_s}$ associated with it. If the size of the massive object is smaller than its Schwarzschild radius, then there will be a *horizon* around the mass located at $\mathrm{R_s}$. It is called a horizon because if a chair or for that matter Shakespeare's collected works fall into the black hole through their gravitational attraction to its mass, then once they pass the horizon, there is no possibility for them to return. Falling into a black hole there is a point of no return. And the points of no return form by definition the horizon. That raises the question what happens to all that stuff disappearing in the black hole. Einstein's theory says without further ado that it disappears in the 'singularity' located at the origin. But that is not what a far away observer sees, because they can not look beyond the horizon. They only see the books approaching the horizon at an ever slower rate. Here a strange complementarity of perspectives arises, because for the infalling 'Hamlet' or 'Midsummer night's dream' nothing special happens as they would smoothly sail though the horizon. From that moment on their fate is decided, they will be swallowed by the singularity; no pardon can be granted, there is just no escape!

So, altogether the physics of black holes was for a long time highly enigmatic, but also unsatisfactory, strangely incomplete and paradoxical to say the least. On the one hand, Einstein's theory inescapably posed their existence, but on the other hand failed to answer many of the basic questions it posed. From the 1970s, fundamental breakthroughs have been achieved in our understanding of black holes. Indeed it turned out that quantum theory had to come in to rescue and resolve some of the bizarre contradictions that black holes confronted the physicists with. Actually it was both quantum theory and information theory that played essential roles. It has become clear that a deep understanding of how black holes work on quantum level could provide the essential keys to a broad understanding and interpretation of what a consistent quantum theory of gravity may ultimately look like.

Black hole thermodynamics

> It is natural to introduce the concept of black hole entropy as the measure of information about the black hole interior.
>
> Jacob Bekenstein (1972)

At first there was the development of a 'thermodynamics of black holes' by Jacob Bekenstein and Stephen Hawking. Subsequently Hawking made the seminal discovery that when quantum processes are taken into account, the black hole is no longer black! Quantum processes that take place at the horizon and which are not allowed by classical physics mean that the black hole will lose energy due to radiation coming of the horizon. Hawking was able to calculate the spectrum of this radiation and that turned out to exactly be the black body spectrum explained by Planck's quantum hypothesis. It means that the mysterious object changed from a black hole into a radiating black body. The black hole is rather like a black ball kept at a certain temperature – appropriately called the *Hawking temperature*, $\mathrm{T_H}$. Let us try to get a grasp of the main components of the argument by recalling some basic concepts:

(i) the thermodynamic relation due to Clausius between heat produced and entropy

$$\mathrm{d}Q = T\mathrm{d}S\,; \tag{I.3.16}$$

(ii) the Boltzmann definition of entropy in terms of the number of states

$$S = k\ln W\,; \tag{I.3.17}$$

(iii) the Schwarzschild radius

$$R_s = \frac{2G_N M}{c^2}\,; \tag{I.3.18}$$

and (iv) the Planck length,

$$l_p^2 = \frac{\hbar G_N}{c^3}\,. \tag{I.3.19}$$

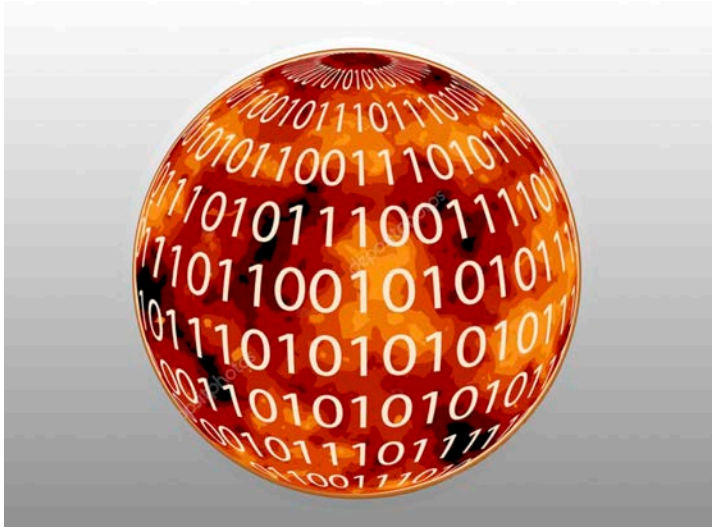


Figure I.3.10: *Information on the horizon.* An artist's impression of bits of information on the horizon of a black hole. The information capacity would be one *bit*, or rather *nat*, per square Planck length.

We start with the observation that classically, nothing can come out of the black hole, so if you drop an object with a certain energy and entropy into the black hole, the only thing you may observe is that the mass increases (and therefore the Schwarzschild radius), but the information content would be lost forever. The idea now is to associate the mass-energy Mc^2 with the heat term in equation (I.3.16) and the area of the horizon A with the entropy term on the right.

Let us talk about spherical black holes, then $A = 4\pi R_s^2$. To convert this area into some entropy, let me define a Planck area a_p, which we will choose as $a_p = 4l_p^2$. The comment here is that the Planck length is the smallest length scale that is physically meaningful, which means that this Planck square is the smallest physically accessible area (as I am giving a heuristic argument I allowed myself to put in the extra factor 4 for convenience). This means that we assume that a single Planck square corresponds to one *nat*

of information[9],

$$S = k\,\frac{A}{4l_p^2} = k\,\frac{Ac^3}{4\,\hbar G_N}\,, \qquad \text{(I.3.20)}$$

which is exactly the expression first written down by Bekenstein and Hawking. In this perspective a black hole would look more like a spherical digital memory as indicated in Figure I.3.10.

This is a surprising result, because, as entropy is associated with the number of degrees of freedom of the system, you would expect that in three dimensions the entropy within a volume bounded by some horizon would grow proportional to the volume and not to the area. This suggests that this rather fictitious, mathematically defined surface will somehow acquire an important physical interpretation if we take quantum processes into account.

Hawking temperature.

> Quantum mechanical effects cause black holes to create and emit particles as if they were hot bodies.
>
> Stephen Hawking (1975)

Next, we want to find the temperature of the black hole as a thermodynamic system. The internal energy is given by $U = Mc^2$ and the entropy S is given by the previous equation. That equation allows us to calculate:

$$\frac{dS}{dM} = \frac{dS}{dR_s}\,\frac{dR_s}{dM} = \frac{8\pi R_s^2}{M}\,\frac{c^3 k}{4\,\hbar G_N}\,, \qquad \text{(I.3.21)}$$

which relates a change in mass with a change in entropy. We may obtain the temperature by using the first law of thermodynamics $dU = dQ - dW$ where the last term on the right-hand side is absent because a black hole doesn't do any work, while for the first term we will use the expression of (I.3.16). This yields the following expression for the internal energy:

$$dU = d(Mc^2) = \frac{\hbar c}{4\pi R_s k}\,dS\,.$$

from which the temperature follows,

$$T_H = \frac{\hbar c}{4\pi R_s k} = \frac{\hbar c^3}{8\pi G_N M k}\,. \qquad \text{(I.3.22)}$$

This is indeed the temperature Hawking derived in his famous paper from 1975, and which was therefore named after him.

We only took the shortest and easiest route which suggests the result, but Hawking really proved that a black hole would radiate as a black stove at that temperature.

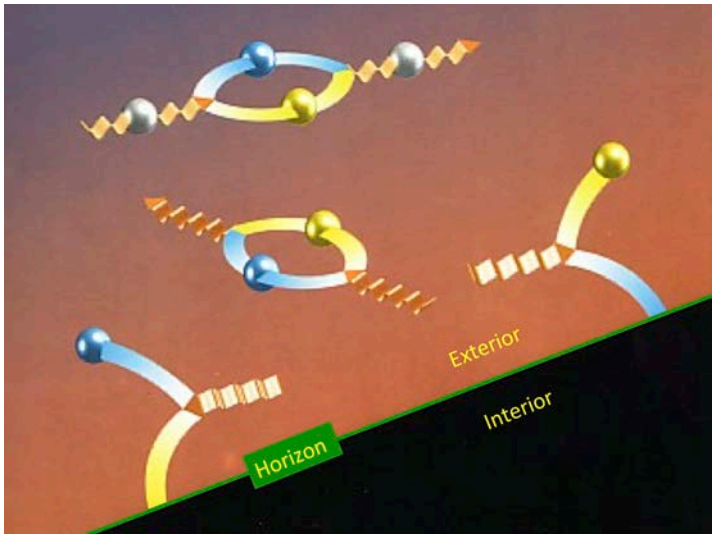


Figure I.3.11: *Pair creation at the horizon.* In a vacuum one always has quantum fluctuations in energy. As a consequence of Heisenberg's uncertainty principle virtual particle anti-particle pairs will be created. Normally these have to recombine but on the horizon there is the possibility that one member of the pair falls in the black hole and the other escapes. This is the microscopic origin of the Hawking radiation.

[9] With N states the information entropy is $H = ^2 \log W$ [bits], the thermodynamical information entropy is defined by $S/k = \ln W$ [nats]. We choose the nat-unit because we want to make the link to the natural logarithm appearing in thermodynamics.

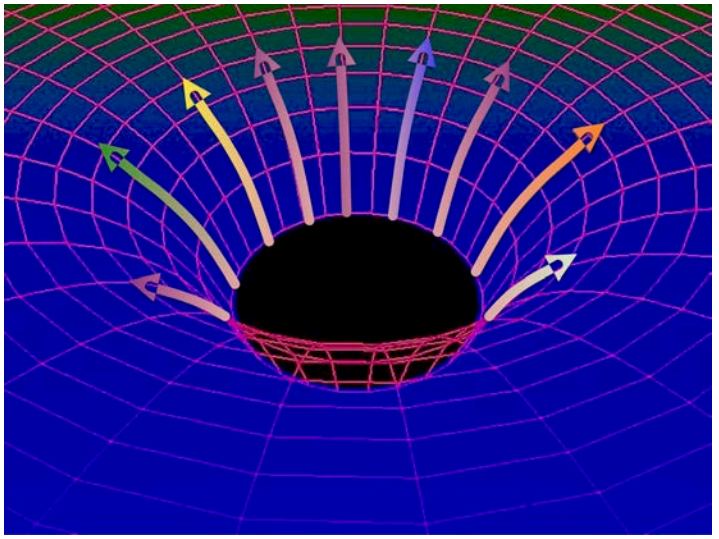


Figure I.3.12: *Hawking radiation.* A black hole is not black as depicted in Figure I.3.4 , but rather as depicted here: a black body at a Hawking temperature T_H, emitting thermal radiation.

The idea was simple but brilliant and the calculation notoriously hard. The starting point of the original derivation was the production of *virtual particle – anti-particle pairs* in the gravitational field near the horizon. He basically calculated the probability that one of the two would fall in, then they could not recombine, and the other particle would have to 'become' real and would be able to escape. The quantum black hole would indeed start radiating and would lose energy. Talking in terms of pictures, we should thus replace the image of the classical black hole of Figure I.3.4 with the quantum version of Figure I.3.12.

Power emitted and life time estimate. The total power emitted by a black body, is given by the radiation law of Stefan Boltzmann (the 'other' Boltzmann) which states that the emitted power would be proportional to the temperature to the fourth power and of course also proportional to the surface area, $P \sim AT^4$, which using equantion (I.3.22) implies that the power emitted would be *inversely* proportional to the second power of its mass $P \sim M^{-2}$. The black hole loses mass because of the Hawking radiation but the more mass it has lost the more it radiates. The final stages are therefore more like an explosion. A black hole would not be black anymore; on the contrary, left on its own in outer space it would *evaporate* until nothing, or may be only some unknown type of remnant, would be left! From the power dependence on the mass, we can make a rough estimate of the life time τ of a black hole, with the following calculation:

$$\int_{t(M_0)}^{t(0)} dt = \int_{M_0}^{0} \left(\frac{dt}{dM}\right) dM \sim M^3 \,. \qquad \text{(I.3.23)}$$

The conclusion is that the life time of a black hole would grow with its mass to the third power.

To put this discussion of black hole evaporation in perspective let us mention that the Hawking temperature for a solar-mass black hole would only be 60 nano Kelvin. Such a black hole presently located somewhere in our universe would absorb far more radiation than it would emit, because the universe itself has at present a background temperature of $2.7K$. This in turn is a consequence of the cooling of the universe by expansion, and is in fact a leftover from the hot Big Bang. So the Hawking radiation phenomena is profound but hypothetical in so far as there is little hope of being able to directly observe it. Consequently, though considered by many to be one of the great discoveries of twentieth century physics, Hawking was not eligible for a Nobel prize.

Surface gravity. The beauty of Hawking's discovery is that it strongly suggests that it is the horizon where the interesting physics of a black hole really takes place. But at the horizon we are still far away from the singularity and space time is smooth. This suggests that we should try to link the emerging temperature of the horizon to a local gravitational concept. The natural candidate would be what is called the 'surface gravity,' which is basically the gravitational acceleration denoted by g at the horizon. We mean the 'universal' gravitational acceleration Galilei was talking about. Indeed, an observer located at the horizon has to

be accelerated to stay there. Just like we are at rest at the earth's surface. The reason we are not freely falling is because the Earth's surface accelerates us radially outward by exerting a normal force. The gravitational acceleration or surface gravity at the horizon is just given by (minus) Newton's law: $g = GM/R_s^2$. Substituting the Schwarzschild radius and comparing with the Hawking temperature we find that the relationship between the Hawking temperature and the surface gravity is strikingly simple:

$$T_H = \frac{\hbar g}{2\pi c k}\,. \tag{I.3.24}$$

It appears that what matters is the transformation from the frame of the observer, freely falling into the black hole for whom nothing special is going on, to the accelerated frame of the observer at rest near the horizon.

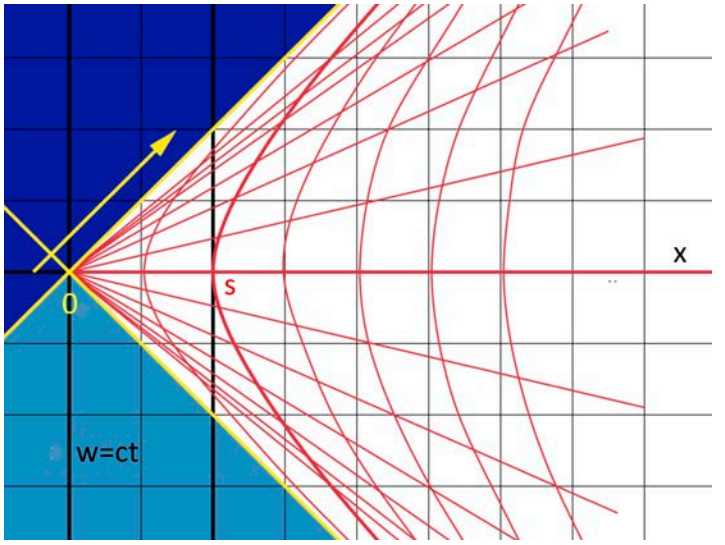


Figure I.3.13: *Accelerated observers.* A space-time diagram for an accelerated frame of reference. The parameter $s = c^2/g$ is the inverse of the acceleration g. The world lines of observers initially at rest on the x-axis are time-like hyperbolae that asymptotically approach the light cones. The white region is the *Rindler space-time* and has a future and past horizon. Light signals emitted from points in the dark region travel along straight lines under 45^o, these do not intersect any world lines and therefore can never be observed by the accelerated observer.

Accelerated observers and the Unruh effect

The above suggests that we should look at the world according to an accelerated observer. This yields another interesting, even more basic link between the structure of space-time and entropy/information known as the *Unruh effect*. Let us first establish that an accelerated observer perceives an horizon. If you transform flat Minkowski space to an accelerated frame you get the so-called Rindler coordinates which is depicted in Figure I.3.13. The world lines of the accelerated observers are time-like hyperbolae. To be precise you should say that the world lines correspond to observers who experience a 'constant force,' because with $F = ma$ and the fact that the mass in this formula is the relativistic mass increasing with the velocity, the effective acceleration becomes smaller so that the velocity never exceeds c, as the figure shows. And that is exactly why the horizon is there. The future light cone of any point beyond the horizon (like the yellow arrow in the dark region) does not intersect with the world line of the accelerated observer and therefore cannot be observed.

This suggests that also in this case including quantum processes à la Hawking will turn that horizon into a black body with a temperature given by the very same formula (I.3.24), linked to the acceleration $g = c^2/s$ of the observer. This remarkable result is known as the Unruh effect, named after William Unruh who first presented the calculations leading to the formula (I.3.24) in 1978. The proof for this case of an accelerated observer amounts to a rather straightforward (quantum) calculation. We start in the rest frame with a quantum field describing some species of scalar particle. The field is in a zero energy state, usually called the vacuum. In this state no particles are present, and that is what the observer at rest perceives. If then we make the transformation to an accelerated frame, the transformed distribution for the density of states corresponds to a ther-

mal energy distribution. This means that the accelerated observer will perceive a highly excited state with many particles present. The spectrum obtained corresponds exactly to the Planck spectrum, the direct consequence of the quantization of energy. This is the spectrum that was hailed as one of the nails in the coffin of classical physics! This calculation by the way beautifully illustrates the relativistic proverb: 'truth is in the eye of the beholder,' and in particular depends on his frame of reference.

Pair creation of charges. In this context it is illuminating to think of the original Hawking argument. Consider a simple two-dimensional $(\mathrm{x.t})$ space-time where one may introduce a constant background electric field, say $\mathrm{E}_0 \approx \mathrm{F}_{01}$ in the positive x-direction. This – in two space-time dimensions – corresponds to a Lorentz invariant background energy. In such a background field there is a certain quantum probability that charged particle anti-particle pairs can be created. Clearly these pairs would split up, the positive charge moving to the right and the negative charge to the left. They would experience a constant force $\mathrm{F} = \pm e\mathrm{E}$ and therefore accelerate, and both would correspond to ideal 'Rindler observers.'

The situation is depicted in the space time diagram of Figure I.3.14 with the two particles accelerating in opposite directions, with their velocities asymptotically approaching the speed of light. They are causally separated from the moment of their creation. The probability of the pair creation depends on the threshold energy which corresponds to the sum of their masses, $2\mathrm{mc}^2$, and the electrostatic energy of the pair depending on their distance d. The remarkable result is that the spectrum corresponds exactly to a thermal distribution matching the Unruh temperature.

One other quantum aspect of profound interest in this example is the fact that in the quantum state in which the pair is created, the particles are entangled. The information of one member of the pair is inaccessible to the other

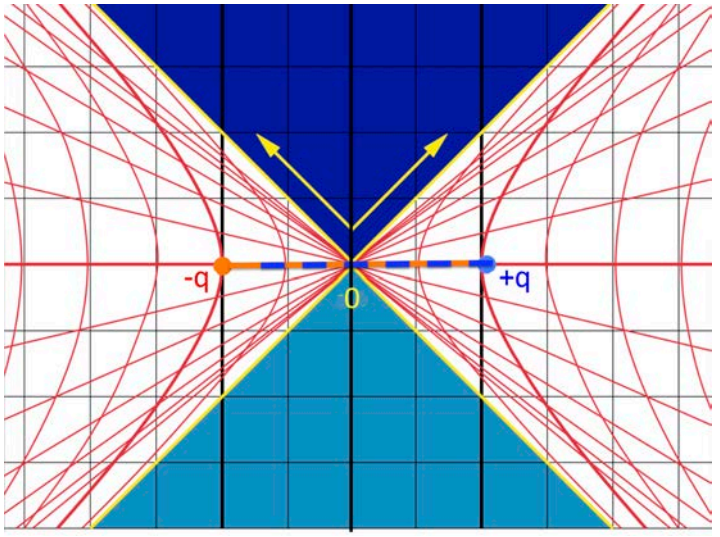


Figure I.3.14: *Pair creation.* In a background electric field, a particle anti-particle pair might be created spontaneously and the members of the pair would accelerate in opposite directions, being causally disconnected from their inception, each living in its own Rindler bubble. However in the quantum world the pair would be entangled which leads to a situation where each of them is dealing with a so-called mixed state.

because it is hidden behind a horizon. This manifests itself in that each of the particles perceives being in a mixed or 'thermal' state with a characteristic entanglement entropy. We will return to these concepts in chapter II.1 but want to mention them already here.

The fate of information. Thermal radiation is completely random (thus maximally uncorrelated and unconstrained) and therefore has maximal entropy. Here we arrive at a familiar point where by solving one question we pose the next one. The idea that quantum processes preserve all entropy and therefore information leads to a non-trivial upgrading of the *information-paradox*: If we throw the entire Encyclopedia Brittanica in a black hole it will be converted into pure thermal radiation according to Hawking. Clearly that cannot be the case, where did all the correlations present in the incoming state go?

If we take a step back, we could compare the formation of a black hole (putting more and more mass on a star, until it becomes a black hole), and its successive evaporation, with a more familiar process (proposed by Sydney Coleman) where we know that quantum processes conserve both energy and entropy. Imagine a piece of coal at zero temperature in a pure state where by definition $S = 0$, that gets irradiated with a fixed amount of high entropy radiation, which we assume is absorbed completely. It brings the coal into an excited state at a finite temperature. As a consequence the piece of coal starts radiating, it will eventually return to the zero temperature state, with zero entropy. As the process of absorbing the initial radiation and emitting the outgoing radiation is a quantum process, it follows that the emitted radiation should have exactly the same entropy as the incoming radiation. And therefore no information could have gotten lost!

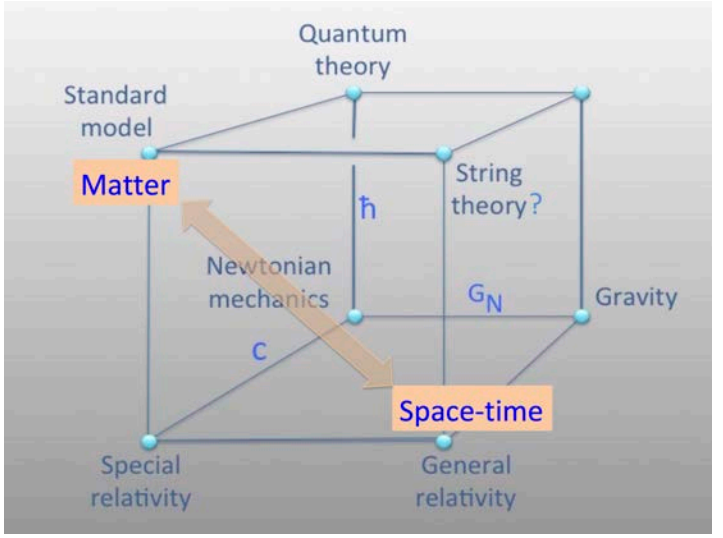


Figure I.3.15: *The magic cube.* Universal constants and the domains of validity of some fundamental theories.

The black hole instability. The conventional narrative is that by throwing an encyclopedia into a black hole, all information would be lost. With the appearance of quantum theory at the horizon, however, our view has radically shifted in the sense that the real physics of black holes is the quantum physics taking place at the horizon. Consequently the question of what happens to the information in the quantum context needs to be critically re-evaluated.

The quantum principles tell you that if you were able to really perform the full quantum calculations including the detailed effects of entanglement, which we haven't discussed yet, then in that case you could in principle recover the entire incoming state. In other words, in the quantum domain it is extremely hard to really get rid of information, it may be hiding, but it still should be out there somewhere. By the way, to people having a Facebook account, this story may sound unpleasantly familiar.

In principle black holes may exist for any mass, hence one may also consider microscopic black holes to rid oneself of the many astrophysical complications that are irrelevant in this context. The statement is that mini-black holes would evaporate very rapidly and therefore be very short-lived. In other words these mini black holes are states of matter that are bound gravitationally, but are unstable, just like many other massive 'bound states' happen to be. This instability gives rise to a finite life time, and the formation and decay process is a quantum process, often referred to as a 'resonance.' In such quantum processes information is preserved, much like the other conservation laws that physics obeys, like the conservation of energy, angular momentum and charge. So, Hawking's crucial discovery has in the end led to a fundamental overhaul of our concept of black holes. And as a consequence the present view is therefore that quantum theory supersedes general relativity in that information has to be preserved somehow. Yet, at this moment, a fully quantum mechanical account of the formation and subsequent evaporation of a basic black hole, which is the litmus test for claiming an understanding of quantum gravity, has not yet been achieved. Though with the advent of string theory as a serious candidate for such a theory, the perspective on black holes has progressed in impressive ways as we will indicate towards the end of

the next chapter. But the fact remains that science at any stage is just 'work in progress.'

The magic cube

The magic cube of turning points. Our story of scientific progress and hope, linked to the successive identification of fundamental constants of nature, is depicted in the 'magic cube' of Figure I.3.15, which is a cube in the space of ideas. The magic cube has classical Newtonian physics on the back lower edge with the laws of mechanics (like $F = ma$) on the left, and his law for gravitation on the right. The constant G_N linking the two appears as the universal constant setting the scale for the strength of the gravitational interaction. The bottom square is the relativity plane where the fundamental constant c (or better $1/c$) is added. The boundary of the domain where Newtonian physics is valid is reached as velocities become of the order of the velocity of light. The Newtonian limit corresponds to $1/c \to 0$. Newton's gravitational force law is instantaneous and therefore incorporates the notion of 'action at a distance'. This notion is incompatible with special relativity, where information and thus disturbances cannot propagate faster than the velocity of light. So special relativity vetoes instantaneous non local interactions. This conflict was then brilliantly resolved by Einstein's theory of gravity, the theory of general relativity.

The vertical dimension opened up with the advent of quantum theory through the universal constant $\hbar$. The classical (non-quantum) limit corresponds to $\hbar \to 0$. The vertical square on the left-hand side includes the modern unified quantum theories for all known forces and matter, except for the gravitational force. The top plane would include a quantum theory of gravity like string theory which so far has not been able to generate predictions that could be tested by experiment. A string theorist may argue that if you had started by postulating string theory, you would have predicted gravity and the other interactions. Such theories unify the notions of matter and radiation with that of space-time. The magic cube illustrates how inconsistencies led the way to fundamental paradigm shifts. Such are the blessings of the inconvenient truths that keep popping up along the winding road of science.

Table I.3.2: Some fundamental sizes and scales.

Notion	Formula	Size [m]
Class. electron radius	$r_e = \frac{e^2}{4\pi\epsilon_0 m_e c^2}$	$\sim 10^{-15}$
Compton wavelength	$\lambda_c = \frac{h}{mc}$	$\sim 10^{-12}$ (e^-)
Strong nuclear scale	$\lambda_\pi = \frac{h}{m_\pi c}$	$\sim 10^{-15}$
Weak nuclear scale	$\lambda_W = \frac{h}{m_W c}$	$\sim 10^{-18}$
Bohr radius	$a_o = \frac{4\pi\varepsilon_0\hbar^2}{m_e e^2}$	$\sim 10^{-10}$
Schwarzschild radius	$R_s = \frac{2G_N m}{c^2}$	$\sim 10^{-2}$ (Earth)
Planck length	$l_p = \sqrt{\frac{\hbar G_N}{c^3}}$	$\sim 10^{-35}$
Age of the universe	$t_0 = \frac{3}{2H_0}$	$\sim 10^{10}$ yr
Thermal wavelength	$\lambda_{th} = \frac{h}{\sqrt{3mkT}}$	$\sim 10^{-11}$ gas 300 K

Conclusion. In this chapter we have celebrated the 'back of the envelope' philosophy and advocated for the virtue of heuristics and approximations. In science the 'truth' is a moving target, elusive like a holy grail because science

is by definition 'work in progress.' And if your work is in progress the notion of truth makes you feel extremely uncomfortable. Every new theory or model is just the next – more sophisticated – working hypothesis. But, as we have shown in this chapter, there is – as every engineer can tell you – a certain pleasure as well as value in playing with the numbers given to you, and applying some dimensional analysis to them. A purist may call it 'recreational physics.' And indeed, that's what we were concerned with so as to get an idea of the relevant scales that are linked to the specific values of *our* 'universal' constants. This game has provided us with some surprisingly deep insights about where quantum effects will rear their heads.

Further reading.
On scales in nature:

- *Mr Tompkins in Paperback*
  George Gamow
  Cambridge University Press,
  Reprint from 1939 and 1944 editions (2012)
- *Knowledge and Wonder*
  Victor F. Weisskopf
  MIT Press (1979)
- *In Praise of Science: Curiosity, Understanding, and Progress*
  Sander Bais
  MIT Press (2010)

On black holes:

- *Gravity's Fatal Attraction: Black Holes in the Universe*
  Mitchell Begelman and Martin Rees
  Cambridge University Press (2020)
- *The Little Book of Black Holes*
  Steven S. Gubser and Frans Pretorius
  Princeton University Press (2017)

Chapter I.4

The quest for basic building blocks

> If, in some cataclysm, all of scientific knowledge were to be destroyed, and only one sentence passed on to the next generations of creatures, what statement would contain the most information in the fewest words? I believe it is the *atomic hypothesis* (or the atomic *fact* if you wish to call it that) that *all things are made of atoms – little particles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another.*
>
> R.P. Feynman (1961)

Figure I.4.1: The human quest for understanding nature.

A splendid race to the bottom

The notion of what the basic building blocks of nature are has repeatedly shifted over time. Every time when a new layer of structure is uncovered a new set of 'basic' building blocks is postulated. That way we turned from chemical elements to atoms, from atoms to the understanding of nuclear structure, and from nuclear structure to the elementary subnuclear particles we know today.

Three levels of simplicity. In Figure I.4.2 we have indicated the subsequent paradigm shifts with respect to the fundamental building blocks of matter. It depicts the frontier of knowledge at three typical moments in the past century, which one could call 'three levels of simplicity.' The first is the level of atoms. The second, nuclear level stands out for its simplicity with only the electron, proton and neutron making up the atoms. The electromagnetic binding of the electrons to the nucleus was provided by the photon while the protons and neutrons were believed to be held together by a nuclear force that was mediated by the *pion*. But this picture is misleading because I left out the 'zoo' of other nuclear particles to be discussed, of which the proton, neutron, and pion are only the most basic and relevant. Finally, at the next level there is the Standard Model of quarks, leptons and force-mediating particles. The fig-

ure provides a bird's eye view of the path of science that brought us ever deeper into matter, and that path is what we are going to run through fast in this section, and explore in more detail in the remainder of this chapter.

The periodic table: atoms. Around 1900 the *chemical elements* were considered the basic building blocks of all matter. Neatly catalogued by the Russian chemist Dmitri Mendeleev in the *periodic table* that he proposed in 1869, the table that has decorated most high-school chemistry-classrooms ever since. These elements are the smallest entities carrying well-defined chemical properties, and as such are indeed the basic building blocks of all of chemistry. The strict order present in the periodic system hinted at an underlying organizing principle that – as we know now – is nothing but the *atomic structure* with a nucleus in the center and electrons 'orbiting' around it, the structure that was uncovered by Ernest Rutherford in 1908 and was so successfully described by the new quantum theory. As we explained in the introduction, quantum physics basically entered our thinking at the atomic level. In this part of the book, and this chapter in particular, we will – as advertised – go down from the atomic level to the physics of the *nuclei* and the underlying structure of elementary *particles.*

Matter matters: nuclei. Once it was realized that the atoms were composite and therefore not truly fundamental, physics turned to the study of the atomic nuclei, which led us to a picture on even smaller scales where we distinguished the *proton* and *neutron* as the building blocks of the nuclei, and of course the electron needed to complete the atoms. And to understand the binding of protons and neutrons in the nucleus a relatively light particle type was identified, the *pion*, that was assumed to be the carrier of the strong nuclear force. It was assumed to play the same role as the photon did for the electromagnetic interactions. Furthermore, it was discovered that the free neutron was in fact unstable; through the so-called β-decay process it would decay into a proton, an electron and another fundamental particle that had to be postulated to save energy and momentum conservation. This elusive particle was called the *neutrino*, a remarkable particle with somewhat ghostlike properties in that it was for a long time believed to have neither mass nor charge, and therefore extremely hard to detect directly.

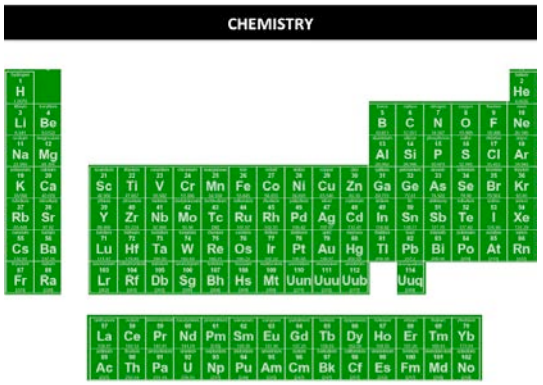
What doesn't meet the eye: the nuclear particle zoo.

> If I could remember the names of all these particles, I'd be a botanist.
>
> *Enrico Fermi*

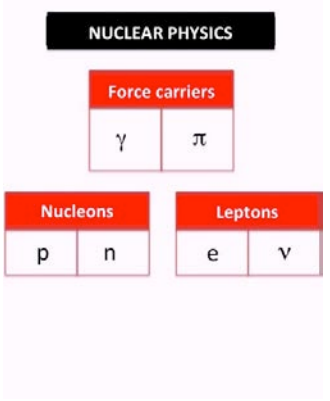
During the 1960s and 1970s experiments demonstrated the existence of an ever-growing list of so-called elementary nuclear particles which was referred to as the *particle zoo*, a term expressing a mild form of despair. Instead of bringing the number of fundamental building blocks back to an ever smaller number, that number seemed to grow without limit. All of these nuclear particles were called *hadrons*. One class consisted of fancy brothers and sisters of the proton and neutron, collectively denoted as *baryons*. A second class contained a large number of relatives of the pions, and those were called *mesons*. Feynman, in one of his popular lectures, quipped that the business of particle physics basically boiled down to a fancy equivalent of smashing watches into a wall, in an attempt to find out what was in them and how they worked.

Law and order regained: the eightfold way. All these new baryons and mesons turned out to be composite as well. It was quite a mess until Murray Gell-Mann (and independently George Zweig) in 1964 created order by applying a beautiful symmetry principle, which Gell-Mann called *the eightfold way*. This term added a spiritual dimension to elementary particle physics as it alluded to the teachings of Buddha, in particular a fragment from the first sermon after his enlightenment, which reads:

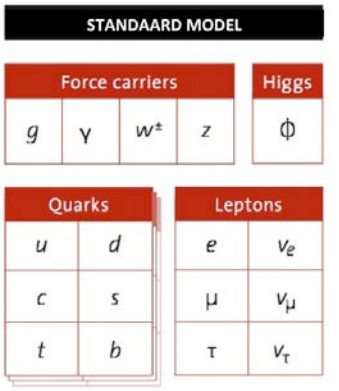
> And what, monks, is the middle path, by which



(a) **Anno 1900.** Mendeleev's iconic periodic table of the chemical elements. The elements are ordered by increasing atomic mass in subsequent lines, and the columns give elements which have similar chemical properties. This structure is a direct consequence of applying quantum mechanics to the atom.



(b) **Anno 1950.** The building blocks of nuclei are the proton and the neutron, and together with the electron they make the atoms. While the electromagnetic force is carried by the photon denoted as γ, the strong nuclear force was believed to be carried by particles called the pion, denoted by $\pi_{\pm}$ and π_0. The neutrino had to be included to account for nuclear $\beta-$decay.



(c) **Anno 2000.** The constituent and force particles of the Standard Model. The quarks and leptons form three families of constituent particles of which only the top row is stable and used to make ordinary matter of the sort listed in the periodic table. Notice that the Higgs particle has a special place in the scheme of things.

Figure I.4.2: *Three levels of 'simplicity.* Three successive levels of reductionism spanning a century of quantum physics. The basic building blocks (a) of chemistry, (b) of nuclear physics and (c) of subnuclear particle physics. The atomic nuclei are built from protons $\mathbf{p}$ and neutrons $\mathbf{n}$ which each consist of three quarks, with $\mathbf{p} = (\mathrm{uud})$ and $\mathbf{n} = (\mathrm{udd})$. So the first element hydrogen $^1\mathrm{H}$ for example has a nucleus consisting of a single proton, while the the second element helium $^4\mathrm{He}$ has a nucleus made up of two protons and two neutrons.

> the one who has thus come has gained enlightenment,
> which produces knowledge and insight,
> and leads to peace, wisdom, enlightenment, and nirvana?
> This is the noble eightfold way, namely,
> right understanding, right intention,
> right speech, right action, right livelihood,
> right attention, right concentration,
> and right meditation.
>
> *Buddha, sermon*

The eight 'rights' mentioned correspond to the corners of the octagon that fits in the big wheel, as shown in Figure I.4.3.

The 'eightfold way' *à la* Gell-Mann is based on a mathematical group of symmetries known as $\mathrm{SU}(3)$.[1] Now, this

[1] $\mathrm{SU}(3)$ is the group of rotations in three-dimensional complex space. Indeed, there is one sentence that always applies to quantum whatever: things become complex! If not in the real sense then at least in the mathematical sense. Numbers, parameters, functions, spaces, transformations, all of it turns complex when you go quantum! You need a tolerance for 'complexification' to avoid *quantum allergy*.

elegant scheme served not only as a meticulous bookkeeping device, and just like Mendeleev's system the eightfold way also made definite predictions for the existence of certain particle types that were discovered subsequently. More importantly, however, was that the $\mathrm{SU}(3)$ structure hinted at the existence of yet a new layer of fundamental particles. Particles from which all known types in the particle zoo could be assembled. Gell-Mann coined the name *quarks* for these new basic building blocks, referring to a – by now famous – quote from the novel *Finnegans Wake* by James Joyce:

Three quarks for Muster Mark!
Sure he hasn't got much of a bark
And sure any he has it's all beside the mark.
But O, Wreneagle Almighty, wouldn't un be a sky of a lark
To see that old buzzard whooping about for uns shirt in the dark
And he hunting round for uns speckled trousers around by
Palmerstown Park?
Hohohoho, moulty Mark!

James Joyce, Finnegans Wake

The pronunciation of this elusive particle's name is 'quork' rather than 'quark', which presumably is the one intended by Joyce as it rhymes with Mark and bark. Irish friends I trust have explained to me that the first exclamation is paraphrasing a typical order in a pub: 'Three quarts (of beer) for Mister Mark!' In German 'quark' refers to a dairy product, and one would interpret it like: 'Three quarks for Master Mark!' It probably is no accident that Gell-Mann in his later life turned to the study of linguistics and in particular to phonetics, in an attempt to trace back the evolution of languages and in some sense reconstruct the 'mother' of all languages. He always had an exceptional fascination and talent for language, as he spoke about twenty of them, and I remember him always taking extreme care to make sure he pronounced the rather unpronounceable names of – in my case, Dutch – colleagues like 'Gerard 't Hooft' or 'Peter van Nieuwenhuizen' perfectly, followed by an exegesis of its meaning and origins!

Figure I.4.3: *The eightfold way.* In Buddhism the 'eightfold way' refers to a very basic principle that brings the eight primary teachings together. It was unfolded in Buddha's first sermon after his enlightenment. Presumably it was the symmetric geometry of the above 'wheel of wisdom' that must have suggested the term to Gell-Mann.

To be or not to be: quarks. According to this scheme the quarks carried a new quantum number which is nowadays called *flavor*. In the original theory there were three 'flavors,' *up, down* and *strange*, denoted by the letters u, d and s. Later on additional flavors were discovered – *charm, top* and *bottom*, denoted by c, t and b – to make a total of six. This would mean that the symmetry group would be the much larger group $\mathrm{SU}(6)$. The fact is that the last three quark types are much heavier particles and very unstable, so they do not play a prominent role in 'ordinary' physics. The physicists say that the $\mathrm{SU}(6)$ flavor symmetry is 'broken' to the much smaller Gell-Mann $\mathrm{SU}(3)$.

The nucleons (and in fact all baryons) consist of three quarks: the proton for example corresponded to (uud) and the neutron to (ddu). The mesons like the pion would consist of quark anti-quark pairs. From this assignment it is not hard to see that these quarks have to carry fractional

electric charges: you have two equations for two charges q_u and q_d, and if you solve them you find that $q_u = 2e/3$ and $q_d = -e/3$.

Splendid unification: the Standard Model. After these new basic building blocks were postulated, it took almost another decade before the real theory for the binding of quarks into the other nuclear particles was developed and the idea of quarks really caught on. What kept the quark idea from general acceptance was the question of 'to be or not to be,' in the sense that these elusive quarks were not observed as free particles. With their fractional charges they would have been easy to identify. For some reason they apparently could not be knocked out of the protons or neutrons. They lived in peaceful coexistence with their 'not being there' so to say. Later we understood that this *confinement* or *imprisonment* property of quarks was a consequence of the nature of the so-called 'color-force' between them. This new fundamental, strong nuclear force between quarks indeed exhibited the desired feature that it imprisons the quarks in threesomes (the *baryons*) or in quark-antiquark pairs (the *mesons*).

It was not until the 1970s that a slow paradigm shift carried us to the Standard Model of quarks and leptons and of the particles that mediate three of the four known fundamental forces. This Standard Model has in the meantime been confirmed in impressive detail by a large number of experiments performed at the major particle accelerators all over the world.

Fatal attraction: forces yield structure

A description of nature does not stop with the inventory of building blocks or basic constituents. One also likes to know why the building blocks stick together the way they do. What we need to know in other words are the forces between the constituents, and how they act. Because it is through interactions between constituents that new structures emerge. This is an all-important ingredient of building models of the world at any level, and we will start with a pedestrian expose, which will deepen along the way in the book. Attractive forces acting between particles may lead to the formation of bound states between the constituents and thus to the formation of structure. Bound systems are only stable when the attractive force is balanced by a repulsive force at small distances. The phenomenon of gravitational binding in Newtonian physics is most familiar. Here we show that our naive classical intuitions fail when talking about the atomic binding of electrons to nuclei caused by the electromagnetic force, and of the nuclear binding of protons and neutrons in the nucleus. We got stuck but quantum mechanics came in to rescue us.

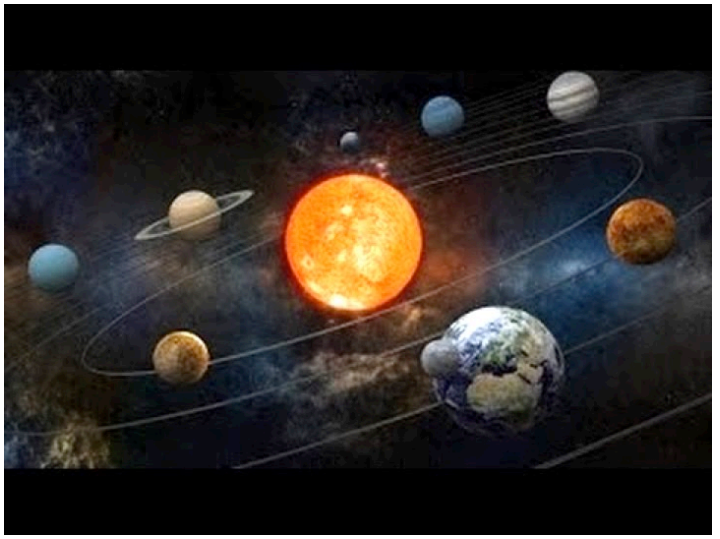


Figure I.4.4: *Gravity at work.* The solar system with its seven (in fact nine) planets moving in bounded elliptic orbits around the Sun. (Source: Getty images)

The Earth orbits the Sun in a slightly elliptic orbit: this binding is caused by the attractive gravitational force as described in the section about Newtonian mechanics. The first question is: why don't we drop into the Sun as the force is attractive all the way in? The force-law corre-

sponds to an infinitely deep gravitational potential well; and so why does the Earth not fall down? Deep wells may provoke deep thoughts. The reason that the Earth doesn't drop in is that it has a tangential velocity, and that velocity induces a outward directed so-called *centrifugal force* that balances the gravitational attraction. More precisely, that tangential velocity component implies that the Earth – Sun system has a certain non-vanishing *angular momentum*, because as you remember $\mathbf{L} = \mathbf{x} \times \mathbf{p}$ and it is the $\mathbf{p}$—component perpendicular to $\mathbf{x}$ that matters. Newton's dynamical laws decree that angular momentum is conserved, basically because the force is directed to the Sun, i.e. in the radial direction, and therefore that force cannot change the tangential component of the velocity.[2] The expression for the energy of a particle in the gravitational field can be written as:

$$E(r) = \frac{p_r^2}{2m} + \frac{L^2}{mr^2} - \frac{G_N m M}{r}\,. \qquad \text{(I.4.1)}$$

The first term contains the radial motion, while the tangential components give rise to the second term, where L is the magnitude of the angular momentum that is a fixed number for each orbiting planet. The last term is the Newtonian gravitational potential. We note that the second term is positive and acts as a repulsive term for decreasing r, while the last is attractive. We have depicted them separately, as well as their sum in Figure I.4.5. The resulting purple curve has a minimum that corresponds to a situation where the radius is fixed and the motion is circular, and the velocity entirely tangential $(p_r = 0)$.

Turned the other way around, one may ask what would happen if we put the Earth at rest at a certain distance from the Sun and let go, then clearly a disaster would be inevitable as the Earth would drop straight into the Sun.

[2] There is something far beyond the scope of our present *exposé* to worry about, however, if we include Einstein's relativity the system would start to radiate gravitationally, which means that the bound system would lose energy and therefore in the end would collapse anyway. This effect of energy loss due to radiation has been observed in spectacular detail in a certain double (neutron) star systems.

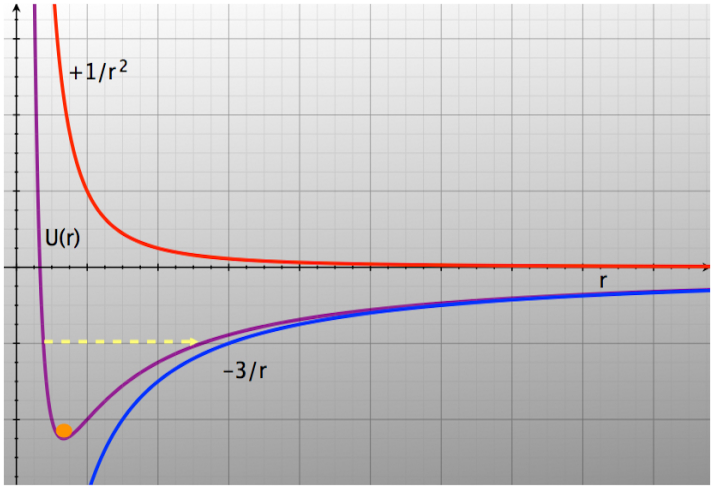


Figure I.4.5: *Balancing attraction and repulsion.* The radial potential $U(r)$ corresponds to the last two terms of equation (I.4.1) and represents a central force field which drops off as the inverse square of the radius. The terms are drawn separately as well as their sum for a particular choice of the parameters. The shape of the potential with a repulsive and an attractive part is universal in situations where we have both bound states with $E < 0$, and scattering states with $E > 0$. The $E = 0$ case represents the parabolic orbit. In the minimal energy (orange dot) the radius is fixed and the motion is circular. If the energy is higher, the orbit can be elliptical (yellow dashed line) with two turning points at different radii.

With $L = 0$ there is no *angular momentum barrier* to save the system from collapse! The potential would correspond to the blue curve in the figure. Well this presents us with a puzzle from a principle point of view, if we take the naive approach and consider the idealized situation where we treat the Earth and Sun as point particles. Then the Earth while approaching the Sun would feel an ever stronger attractive force giving the Earth an ever-growing acceleration and speed! And by falling in, the Earth would gain an 'infinite' amount of energy, and as its speed would be limited by the speed of light it would acquire an unlimited amount of mass.

What actually happens in such radial approaches may cer-

tainly be violent as we know from falling meteorites hitting us from time to time, but due to the finite size of the objects colliding the acceleration towards the center is stopped and the kinetic energy is converted into structural damage, debris flying around, and heat.

Infinities call for new physics. What these examples teach us is that other, non-gravitational physics takes over and saves the day. This is often the cavalier way physicists wave their hands about the singularities in their theories that keep pestering them, and that most non-physicist audiences are most curious about. It must be said that the physicists have been unreasonably successful with this pragmatic approach. As far as we know nature *is* non-singular, and the moment it threatens to become singular, it usually amounts to a wake-up call to go and search for new physics and new theories that avoid the singularities and thereby save serious science from demise.

This is exemplified by applying the gravitational force to the different case of a radially collapsing star. If we look at an extremely massive star, the gravitational attraction is directed to the center and is kept in balance by the repulsive force, caused by the outwardly directed pressure generated by the nuclear burning processes in its core. However, as we'll discuss later on, the amount of nuclear fuel is finite and even a massive star will one day stop shining, after which a gravitational collapse to some compact object is unavoidable. Depending on the mass of the original star, this final state can be a white dwarf, a neutron star or a black hole. In the first two cases a new repulsive force working at smaller inter-particle distances halts the collapse and allows for a new balance thereby avoiding the singularity. The most dramatic possibility is the formation of a black hole. But a black hole is surrounded by a horizon that keeps us from knowing what happens to the mass inside and whether there is anything singular going on. A horizon seems to save the day, or better the horizon masks our ignorance about what precisely is going on! Putting things behind the horizon sounds like the scientific equivalent of sweeping things under the carpet. Yet, that is apparently the way in which nature prefers to keep some of its secrets. This property is referred to as *Cosmic censorship*.

In the previous chapter we mentioned the direction in which progress is made to handle this problem. It is again by shifting the attention from the singularity in the origin to a deeper quantum mechanical understanding of what a horizon really is. In principle black holes come in all sizes and a Planck-mass black hole would have a horizon as well, and could therefore be considered as the 'hydrogen atom' of quantum gravity. We just don't know yet how this works precisely, as we have no fully consistent quantum theory of the gravitational force. But taking the essential idea of Hawking radiation from the horizon as a guiding principle, black holes would be unstable states of matter, bound to somehow evaporate completely. And that would turn the embarrassment of its singularity in some kind of red herring. For the moment however, black holes remain in the category of 'unsolved problems'.

The quantum stability of matter. In the case of colliding ordinary objects it is the much stronger electric force that keeps the balance, and prohibits the infinite energy gain of two point particles colliding gravitationally. But what if we have two point particles with opposite charges, say a positively charged proton and a negatively charged electron, which make up the familiar hydrogen atom? Now both forces are attractive, and yes there can again be an angular momentum barrier, or better a repulsive core due to the angular momentum that dominates over the attraction for small distances. But what about the lowest state where the angular momentum would be zero.

Classically the same 1/r singularity – as it is called – would certainly rear its head again, and maybe you would expect a mini-blackhole to form. No, this is certainly not what happens, and yes, there is other physics – quantum physics to be precise – that saves the day. The lowest quantum

state with zero angular momentum turns out to be perfectly stable and well behaved. It has a wavefunction that corresponds to a spherically symmetric probability distribution for the electron to be at a finite distance from the nuclear core. It is one of those 'life saving' manifestations of the *Heisenberg uncertainty relation*. This relation does not allow a quantum particle to just sit at the bottom of a 'quantum bowl'; being at rest and completely localized is a no-go. Heisenberg prohibits a particle from falling down to the origin. This result is all-important because what it means is that quantum theory guarantees atomic stability. Stability means that the energy of a quantum system is somehow bounded from below. The atom can radiate away electromagnetic energy by emitting photons until it gets down to a lowest angular momentum and lowest energy state which is perfectly regular and stable.

Having made this victorious claim I should sit back for a moment and scratch my head. What about an atom with more than one electron? Just take any. Would this atom not decay into a state where all electrons descend to their lowest possible, so-called, ground state, one may ask? Certainly if we ignore the electric forces between electrons. But is this what we see happening?

The answer to this well-posed question is a fully-fledged 'No'! We see that different atoms behave quite differently from a chemical point of view, and that fact is at the root of all diversity in nature. How could that ever be if all electrons would be sitting in the same state? This disturbing shortcoming of naive quantum theory is resolved by an additional – at first sight magical – quantessential principle, that prohibits particles like electrons to occupy the same state! When Wolfgang Pauli introduced this *exclusion principle* it was certainly a rather ad hoc rule, a veritable *deus ex machina*. But it did in one blow bring theory back into excellent agreement with the observations. According to this principle you should think of electrons a bit like people at a pop festival in desperate need of a toilet. The simple truth is that a 'seat' is either free or occupied and there is no in-between; if occupied, you have to go and look for the nearest free seat, which may be way out. Electrons are permanently involved in playing some game like 'musical chairs.' A notable aspect of this mutual exclusion is that it only concerns exclusion of the same type of particles, not particles of a different type. Moreover not all particle types are subject to the exclusion principle. The particles which are like electrons are called *fermions*, while the particles that are not, like the photon, are called *bosons*. We will return to this topic in a forthcoming section. First we turn to a more detailed description of the atom.

Atomic structure

One of the early icons of quantum theory is the Bohr model of the atom that we discussed in the previous chapter. It makes it clear in a transparent way how a rather simple but radical idea that can be directly implemented leads to a very non-classical behavior, explaining qualitatively the physics we are observing. This heuristic device was then turned into a mathematical precise framework by Heisenberg, Schrödinger, Dirac, Born and many others. This work revealed a complete set of quantum numbers labeling the states including the spin of the electrons. To complete the model of the atom Pauli's exclusion principle also had to be invoked The study of the atom taught us what quantization really means, and at the same time raised the intricate epistemological questions that haunted the theory and its practitioners for almost a century thereafter.

The Bohr atom: energy quantization

In the subsection on the Bohr-radius on page 134 of the the previous chapter, we introduced the Bohr model for the atom with its characteristic quantized orbits depicted in Figure I.3.7, and its quantized energy levels. In this sec-

tion we want to look at these quantized energy levels and point out how they are related to the observed discrete line spectra of light emitted by atoms. The connection that Bohr established was that if an atom makes a transition from some excited state to a lower one, it would emit a photon with a frequency given by the Planck – Einstein relation, so $\Delta E = h\nu$. Conversely, an atom could absorb a photon if its frequency matched the energy for an electron to move up. The schematic of such processes is given in Figure I.4.6, where it is also indicated that for the hydrogen atom the transitions to the ground states have frequencies that correspond to the ultraviolet, while the transitions to $n = 3$ correspond to the infrared end of the spectrum. So only the transitions to the $n = 2$ levels are in the visible domain. Clearly having a simple model that could account for these discrete line spectra was a major success for the early quantum physicists. These line spectra can be considered as an atomic barcode, if you hand it to me I can tell you which atom you were looking at.

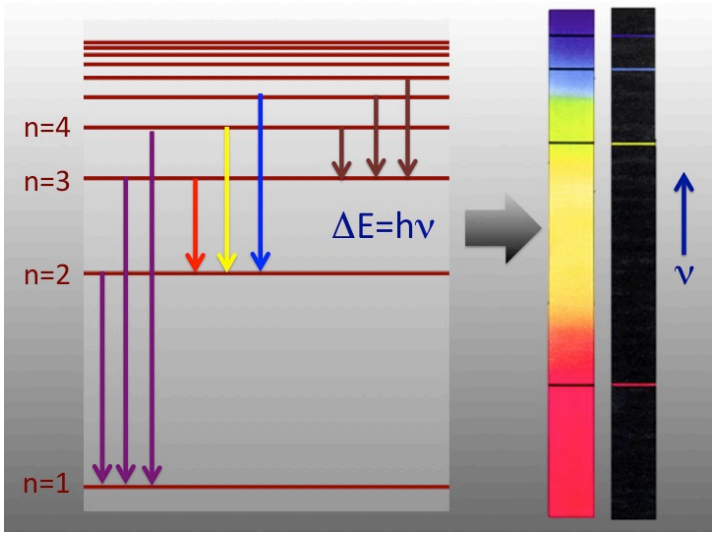


Figure I.4.6: *The origin of light.* If the electron makes a transition between the energy levels, the fixed energy difference ΔE translates into the photon frequency; $\Delta E = h\nu$. This determines the color of the lines of the spectrum, which can be observed in absorption (left) or in emission (right).

.

The Schrödinger atom: three numbers

After the Bohr model was introduced in 1913, it would take another thirteen years until Schrödinger and Heisenberg published their fundamental equations for quantum physics. The first called the theory *wave mechanics* and the second *matrix mechanics*, but in fact they were fully equivalent descriptions of the quantum states and their observables, as was later shown by Dirac. The Schrödinger equation is a wave equation in three dimensions, that could be solved exactly for simple atoms and that yielded the full spectrum of atomic states with all its quantum numbers. It went much further than the Bohr model, but to a certain extent it incorporated the same simple idea in a full three-dimensional model for the atom. In the Schrödinger picture the states correspond to wavefunctions $\psi(x)$ that are defined over all of the position space, $x \in \mathcal{X} = \mathbb{R}^3$. And from the wavefunction of a state the related probability distribution of where to find the electron can be derived.

The equation: a guided tour. 

So, let me step back and try to give you an idea what the Schrödinger equation is about, and what it looks like. Let us call it a 'guided tour.' In an operational, maybe even opportunistic, sense, going from classical to quantum mechanics, is mathematically speaking not that hard. Once you accept that momentum is represented as a spatial derivative operator, $\mathbf{p} = -i\hbar\boldsymbol{\nabla}$, and the energy or Hamiltonian as a time derivative $H = i\hbar d/dt$, one can translate the classical functions into corresponding quantum operators or equations just by substitution. For example:

$$E = \frac{p^2}{2m} + V(x) \quad \rightarrow \quad i\hbar\frac{d}{dt} = -\frac{\hbar^2}{2m}\boldsymbol{\nabla}^2 + V(x)\,, \quad \text{(I.4.2)}$$

on the left we have the Newtonian expression of the energy and on the right we have the Schrödinger 'wave opera-

tor', which when we let it work on a (wave) function $\psi(x,t)$ yields the *Schrödinger equation* in all its glory:

$$i\hbar\frac{d\,\Psi(x,t)}{dt} = \big(-\frac{\hbar^2}{2m}\boldsymbol{\nabla}^2 + V(x)\big)\,\Psi(x,t)\,. \qquad \text{(I.4.3)}$$

For now we don't want to get too deep into the mathematics of the equation but let us at least make some observations which are not so hard to digest:
(i) The equation expresses a simple truth, namely that the energy (operator) generates the time evolution of the system.
(ii) Quantum states are described by wavefunctions Ψ that satisfy this equation.
(iii) The wavefunction is complex meaning that it has a phase factor in it, and it describes a probability amplitude.
(iv) Squaring the amplitude gives the probability density $p(\mathbf{x},t)$ for finding the electron in a small volume element $d^3\mathbf{x}$ around the position x and at a time t. We defined $p(\mathbf{x},t) = |\Psi|^2$, so that the overall phase of the amplitude drops out. It doesn't affect the probability, which is where the physics is.
(v) Indeed, the notion of probability apparently enters already on this basic level in the theory, where we are still talking about the state of a single particle.

The quantization. Of great importance are the so-called *stationary states*, meaning that the physical properties do not change in time. You would think that the wavefunction has to be time independent in that case but that is not quite true. What is true is that the time dependence has to sit in the phase factor $\phi(t)$ which is going to drop out anyway in the probability density. The answer is to write Ψ as a product of a phase factor which depends on t only, and a time independent wavefunction $\psi(\mathbf{x})$ that describes a time independent stationary state. We write

$$\Psi(\mathbf{x},t) = \phi(t)\psi(\mathbf{x}) = e^{-iEt/\hbar}\,\psi(\mathbf{x})\,, \qquad \text{(I.4.4)}$$

and substitute it in the Schrödinger equation. If you take the derivative, you get that the time dependence drops out completely and you are left with a nice time independent equation for $\psi(x)$:

$$\big(-\frac{\hbar^2}{2m}\boldsymbol{\nabla}^2 + V(\mathbf{x})\big)\,\psi(x) = E\,\psi(x)\,. \qquad \text{(I.4.5)}$$

where E is the constant energy value of the stationary state ψ. The crucial point here is that you *first* have to solve the equation to find out which values of E make quantum sense. It turns out that only specific values give a solution for which the square of ψ gives an acceptable probability function. This means that the solutions have to be *square integrable*; the integral over all of space of the absolute square of the function has to be finite. This integral can then be normalized to one to obtain an appropriate probability density. This type of mathematical problem is called an *eigenvalue problem*; the values E that occur in equation (I.4.3) are called *eigenvalues* and the corresponding functions $\psi(\mathbf{x})$ are called *eigenfunctions*. This really is the stage at which the *quantization* 'takes place' in the Schrödinger approach, and the eigenvalues are often called *quantum numbers*. Hopefully this helps you to also imagine what people mean when they talk about 'quantizing' some (classical) system. They perform the substitutions as we did in equation (I.4.2) and then look for the eigenvalues and the corresponding eigenfunctions characterizing the quantum states of the system. ■

A free quantum particle. Let us consider the simple case where $V(\mathbf{x}) = 0$ that corresponds to a free particle. The solutions are periodic plane waves:

$$\psi_{\mathbf{k}}(\mathbf{x}) \simeq e^{i\mathbf{k}\cdot\mathbf{x}}\,. \qquad \text{(I.4.6)}$$

The meaning of the vector $\mathbf{k}$ (which appears here as a vector of free parameters defining the solution) becomes clear if we substitute the solution in equation (I.4.5) with $V = 0$, which yields:

$$E_{\mathbf{k}} = \frac{\hbar^2|\mathbf{k}|^2}{2m}\,. \qquad \text{(I.4.7)}$$

This is just the classical expression for the kinetic energy once we use the fact that the momentum is given by $\mathbf{p} =$

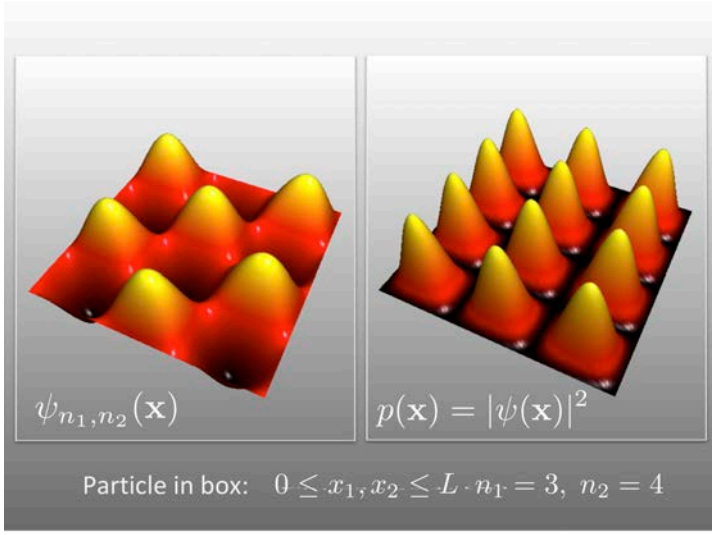


Figure I.4.7: *Quantum particle in a box.* A state of a two-dimensional quantum particle in a box of length L. The wavefunctions $\psi(\mathbf{x})$ have to vanish on the boundary, and are of the form

$$\psi_{n_1 n_2} \sim \sin(n_1 x_1 \pi/L) \sin(n_2 x_2 \pi/L)\,.$$

We have plotted the wavefunction ψ and corresponding probability density p for finding the particle corresponding to quantum numbers $n_1 = 3$ and $n_2 = 4$.

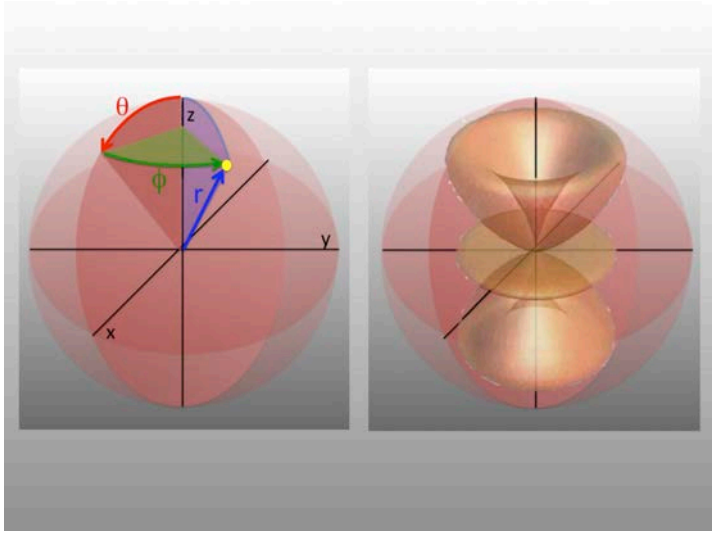


Figure I.4.8: *Spherical harmonics.* The spherical coordinates r, θ, and φ of the (yellow) point $\mathbf{x}$ are defined on the left. On the right the angular distribution $\rho_{l,m}(\theta,\varphi) = \left|Y_l^{\,m}\right|^2$, for a state with quantum numbers $l = 3,\ m = 1$ is plotted.

$\hbar\mathbf{k}$. There is an annoying technical complication here, if you calculate the probability density for the particle, you find $p(x) = |\psi|^2 = 1$, which is unacceptable because it cannot be normalized to '1'. If you take the integral over over a constant non-zero probability density then you would find the total probability to be be infinite! The way out is to put the particle in a box, say a cube of size L, so that the wavefunctions have to vanish on the boundary where $x_i = L$. This in turn means that the momentum values become quantized: $p_i = \hbar k_i = \pi\hbar n_i/L$ with integer-valued n_i. The space of admissible momenta corresponds therefore to an infinite three-dimensional cubic lattice, where the energy levels grow as the length of the momentum vector squared: $E_{\mathbf{n}} \sim \mathbf{n}^2$.

In Figure I.4.7 we have depicted a particular solution for a two-dimensional particle in a box, where the wavefunctions that satisfy the boundary conditions are of the form:

$$\psi_{n_1 n_2}(\mathbf{x}) = N \sin(n_1 x_1 \pi/L) \sin(n_2 x_2 \pi/L)\,, \qquad \text{(I.4.8)}$$

with N a normalization factor. In the figure we plotted the wavefunction and the corresponding probability density function for the case $n_1 = 3$, $n_2 = 4$. Note that this wavefunction describes a one-particle state, but that that particle has a rather outspoken preference for certain positions which sit on a periodic lattice inside the box. We will in Volume II go much deeper into what this probability interpretation of the wavefunction exactly means. For example, looking at the figure the obvious question: 'Where is the particle?' begs for an answer. As it turns out that answer is far from obvious!

The hydrogen atom. Let us return to the question of what the states look like for an atom. With the nucleus in the origin the electron moves in the spherical Coulomb field caused by the positive charge of the nucleus. The potential has a rotational symmetry, which means that it is ad-

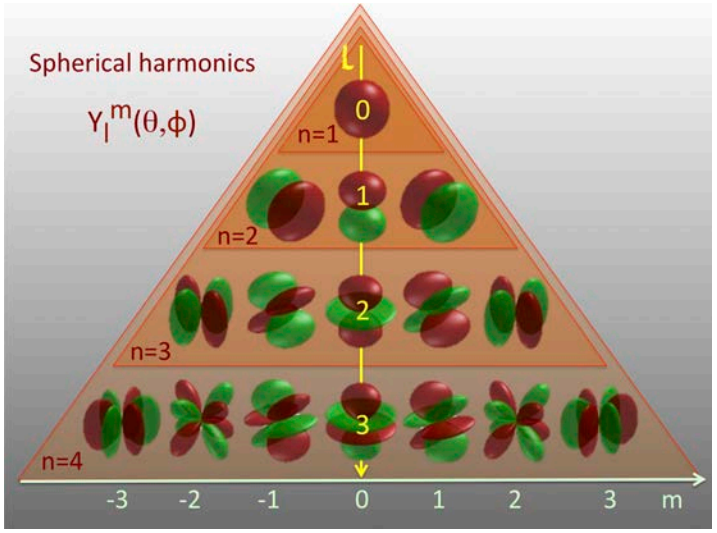


Figure I.4.9: *What the quantum states of hydrogen look like.* The angular dependencies of the hydrogen wavefunctions corresponding to the so-called spherical harmonics $Y_l{}^m(\theta,\varphi)$ where n, l and m are the discrete quantum numbers which label the state. Each state can hold at most two electrons, one with spin up and one with spin pointing down. The first quantum number n labels the energy level corresponding with a triangle in the Figure. At each level we have states where angular momentum l runs from 0 to $n-1$ along the vertical axis and for each l the component m runs horizontally from $-l$ to l.

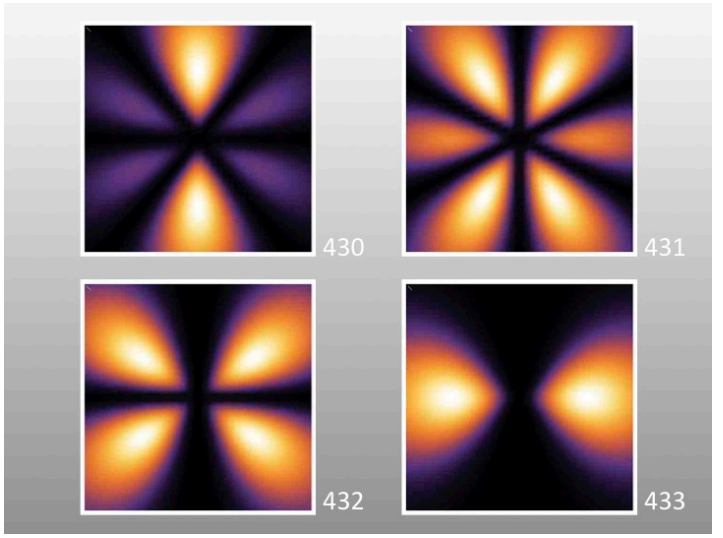


Figure I.4.10: *Charge distributions.* Light color indicates high probability.The charge or electron probability distribution in the $xz-$plane shows the θ and r dependence. Depicted are the $n=4, l=3$ states, with $m=0,...,3$. These states correspond to the states on the bottom line of the previous figure. The shapes of the probability distributions are all-important for understanding the chemical binding properties.

vantageous to rewrite the Schrödinger equation in terms of a radial (r) coordinate and two angular coordinates (θ,φ) (see Figure I.4.8). The equation then basically separates into three independent equations depending again on certain discrete quantum numbers. The radial quantum number $n=1,2,\ldots$ linked to energy level is basically the orbital quantum number introduced by Bohr. And the energy eigenvalues E we just discussed are quantized like $E\sim 1/n^2$. The angular dependence of the states introduces two more quantum numbers: $l=0,\ldots,n-1$ and $-l\leq m\leq +l$, both of which are related to angular momentum. The wavefunctions corresponding to the states are usually written like $\psi_{nlm}(r,\theta,\varphi)=R_{nl}(r)Y_{lm}(\theta,\varphi)$ where the radial and the angular dependences are separated. In Figure I.4.9 we have depicted the angular dependencies by plotting (the real part) of the functions $Y_l{}^m$ for all admissible l and m values up to principle quantum number $n=4$.

Degeneracies. It turns out that the states $\psi_{nlm}(x)$ are highly degenerate, meaning that different states will have the same energy. For every energy level (labeled by the quantum number n) there are a total of $2n^2$ different angular momentum states which all have the same energy. In Figure I.4.9 these degenerate states correspond to the angular (l,m)-states within each triangle labeled by n. The extra factor two comes from the two possible electron spin states that will be discussed shortly. Plotting this discrete spectrum one would get a three-dimensional discrete lattice filling a triangular pyramid (or is it a nicely decorated Christmas tree?). These degeneracies are not accidental: they are the consequence of certain symmetries in

this problem. These symmetries lead to certain conserved quantities and these in turn lead to degeneracies in the energy spectrum. We will return in more detail to this topic in Chapter II.6.

Lifting degeneracies. These degeneracies corresponded exactly to the observations that the Dutch physicist Pieter Zeeman made some 25 years earlier (almost simultaneous with Planck's quantization hypothesis). He discovered that by turning on a magnetic field the degeneracy of the different angular momentum eigenstates was lifted, which is reflected in the splitting of the spectral lines corresponding to one energy level into many different lines. So you could count the multiplicity of the degeneracies. For the discovery of this 'Zeeman splitting' he shared the Nobel prize with Hendrik Antoon Lorentz in 1902.

With the Bohr model in mind it is intuitively not too hard to interpret these splittings. Clearly Bohr had only used circular orbits but if we think of negatively charged electrons orbiting the positively charged nucleus, these would create a magnetic moment like a circular electric current would do. This magnetic moment would be proportional to the angular momentum of the electron state. What caused the Zeeman effect was that the different magnetic moment or angular momentum states would acquire an extra energy contribution from the interaction of that moment with the external magnetic field. And interpreted this way, his measurements showed direct evidence for the quantization of the component of the angular momentum along the magnetic field in integer multiples of $\mathrm{m}\hbar$, where for given l there was naturally the restriction $-\mathrm{l} \leq \mathrm{m} \leq +\mathrm{l}$. Needless to say that none of these quantization rules can be understood from a classical point of view.

This splitting, which could be completely accounted for within the framework of the Schrödinger or Heisenberg equation, is called the *normal* Zeeman effect. However, Zeeman did actually discover an additional quantessential feature in the spectra, which is referred to as the *anomalous Zeeman effect*, to which we turn next.

The discovery of spin

> The Pauli principle was published early in 1925. .... Well, I had introduced those quantum numbers but, if I had been a good physicist, then I would have noticed already in May 1925 that this implied that the electron possessed spin. But I was not a good physicist and thus I did not realize this... Then Uhlenbeck appears on the scene ... he asked all those questions I had never asked ... When the day came that I had to tell Uhlenbeck about the Pauli principle – of course using my own quantum numbers – then he said to me: 'But don't you see what this implies? It means that there is a fourth degree of freedom for the electron. It means that the electron has a spin, that it rotates'... I asked him: 'What is a degree of freedom?' In any case, when he made his remark, it was luck that I knew all these things about the spectra, and I said: 'That fits precisely in our hydrogen scheme which we wrote about four weeks ago. If one now allows the electron to be magnetic with the appropriate magnetic moment, then one can understand all those complicated Zeeman – effects.'
>
> *Samuel Goudsmit (1971)*

As announced, there was another quantessential treasure hidden in Zeeman's spectral data that caused a great deal of confusion among the early quantum physicists. It is known as the *anomalous Zeeman effect*, and was observed in the spectrum of Sodium, where a line in the absence of an external magnetic field already appeared split: this is because of the coupling between the spin and orbit magnetic moments. When Zeeman turned on the field, he found further splittings in an even number of lines as indicated in Figure I.4.11. These splittings implied that quan-

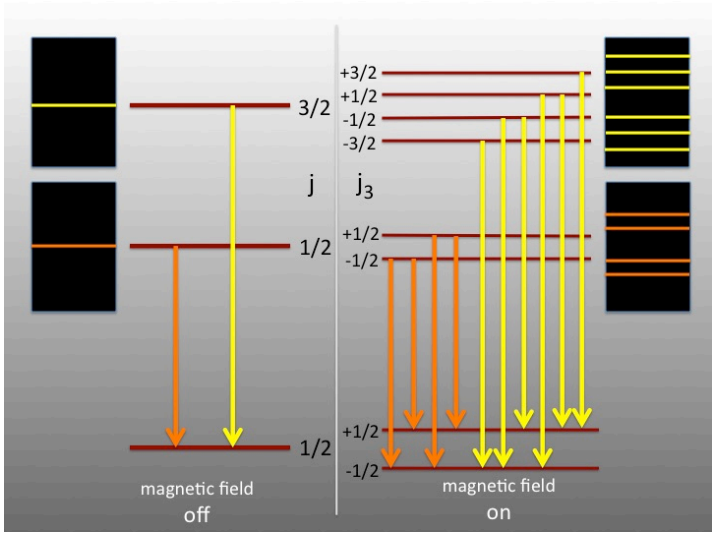


Figure I.4.11: *The discovery of* spin *(and the* qubit*).* This shows the anomalous Zeeman effect discovered in 1898, the same year that Planck introduced his constant. The line spectrum of Sodium, corresponding to the arrows in the figure, was split due to the spin of $1/2$ of the electron. Without the magnetic field, the level l split up into levels with total spin $\mathrm{j} = \mathrm{l} \pm \frac{1}{2}$ (on the left). Turning on a weak magnetic field he observed the *hyperfine structure* because the spin degeneracy was lifted.

tum theory would somehow also admit half-integral values for the angular momentum (as $2\mathrm{j}+1$ is only even if j is half-integral).

In 1925 the resolution was proposed by two young Dutchmen, Samuel Goudsmit and George Uhlenbeck who were still graduate students at Leiden University. They came up with the bold proposal that the electron was spinning around its axis and that that 'spin' would account for the so-called hyperfine splittings observed in the anomalous Zeeman effect. It may remind you of the good old solar system with the Earth rotating about its axis while orbiting the Sun! The idea implied that the observations had nothing to do with an extra feature of the atom as a whole, but rather with a totally new feature of the electron itself.

Behind the scenes. Wolfgang Pauli had already come across this problem in 1925 and had understood that the quantum numbers of atomic states were basically related to the radial and the angular motions of the electron. Indeed, three dimensions gave rise to three quantum numbers: $\mathrm{n} = 1, 2, \ldots$ for the radial direction, and $\mathrm{l} = 0, 1, \ldots, \mathrm{n}-1$ and $-\mathrm{l} \leq \mathrm{m} \leq +\mathrm{l}$ for the angular motions. But he also noted that to get things right he needed a fourth quantum number which he somewhat desperately called *Zweideutigkeit*, meaning something like 'double valuedness'. The story of how the all-important discovery of spin unfolded is a kind of amusing, but for the young researchers involved in fact rather traumatic.
Goudsmit and Uhlenbeck, discussed their spin-idea with their Leiden advisor Paul Ehrenfest, who liked it and proposed that they should write it up. They did so and showed their work to the grand old Leiden professor Lorentz who had earlier developed a sophisticated theory of the electron, but entirely within the classical framework. A thing he could do well was to calculate the rotational speed the electron would need to have in order to produce the magnetic moment corresponding to $(1/2)\hbar$, and that turned out to exceed the speed of light by orders of magnitude. This is in clear contradiction with the theory of relativity. Understandably, this argument knocked down the confidence of the students and they went back to Ehrenfest to humbly withdraw their paper that contained this incredible stupidity. Alas, it turned out that Ehrenfest had already submitted the paper, and didn't seem to take it too seriously, making the consoling remark: *'Sie beiden sind jung genug sich eine Dummheit leisten zu können.'* ('The two of you are still young enough that you can afford yourself such a stupidity'). Actually it

seems that Bohr when he heard about the proposal liked it and Einstein also apparently judged it rather mildly.
It actually turned out that somewhat before that time, a young American physicist, Ralph de Laer Kronig, had also thought of the electron spin (amusingly, I took my first quantum mechanics course with Kronig in Delft in 1966). To his misfortune, he happened to show the idea to Wolfgang Pauli, who instantly demolished it, so that Kronig ended up not working on it any further. The issue of who should and who should not be credited with the discovery/invention of spin remains hidden in darkness. It is this strange story with a touch of tragedy that may explain why a Nobel prize was never awarded for the discovery of the quantessential property of spin as an intrinsic property of particles. It also shows that the advice of even the greatest 'advisors' should sometimes be taken with the necessary pinch of salt. □

The electron possessed a new property called *spin*! It could only exist in a spinning state with intrinsic angular momentum values $s = 1/2$ in units of $\hbar$. In Figure I.4.11 we show how this conjecture did in a rather spectacular way resolve the special properties of those particular 'D-lines' in the spectrum of Sodium. The idea was to think of a new *total* angular momentum quantum number denoted $j = l \pm 1/2$, basically expressing that the spin would either be aligned or anti-aligned with the orbital angular momentum. In that case the component along the field of j denoted as j_3 could run from $-j \leq j_3 \leq +j$. Hence the $2j + 1$ energy levels of the right-hand side of Figure I.4.11 is an even number since $j = 1/2$ and $3/2$ respectively. And that does the job if you assumed in addition that the transitions could only take place if they obeyed the rule $\Delta j_3 = -1, 0, 1$, that followed naturally if you took into account that the outgoing photon itself had spin one.

Let us conclude with a comment on the splittings of the energy levels. If we would have refined our model to include the interaction of the magnetic electron spin degree of freedom with the magnetic moment due to the orbital motion of the electron, the so-called *spin-obit coupling*, we would have found the *fine splitting* of the left column in Figure I.4.11. Furthermore if we would have included the interaction of the electron spin with the nuclear magnetic moment, we would have found the hyperfine splittings, on the right of the figure.

Fermions and bosons

There are many macroscopic phenomena that can only be understood from underlying quantum principles of matter. One of the quantum principles which has a tremendous explanatory power is Wolfgang Pauli's exclusion principle*: it decrees that two electrons cannot occupy the same quantum state. This exclusion property is instrumental, for example for understanding the atomic structure of the elements and the magnificent chemical diversity that derives from it. Not all particles obey the principle though: the particles that have half-integral spin do obey and are called* fermions*, while the particles that have integral spin do not and are called* bosons*.*

Having made a strong plea for the microscopic domain as the realm where the laws of quantum theory are indispensable, I should hasten to correct myself. This is a severe understatement. Quantum theory manifests itself on all scales, but could only be discovered on the microscopic level where it is omnipresent, manifest and inescapable. Once that is recognized, however, it turns out that there is a host of macroscopic phenomena that cannot be explained without a deep understanding of quantum theory. This is so because macroscopic systems are made up of large numbers of microscopic quantum particles. One might expect that there are particular proper-

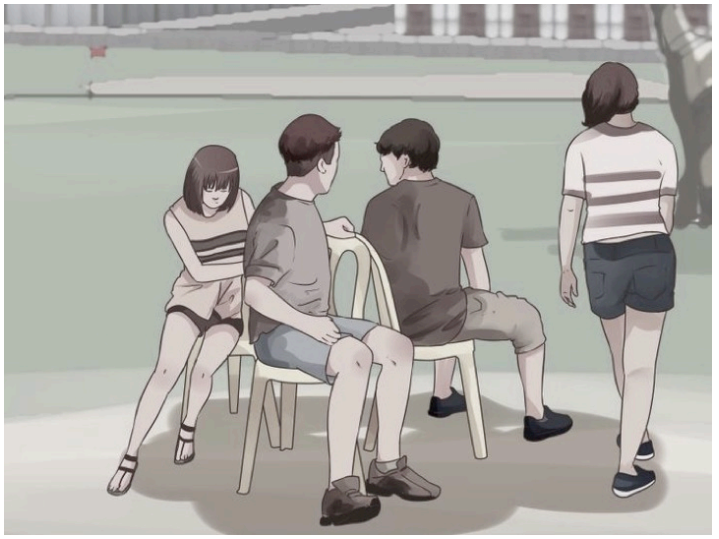


Figure I.4.12: *The exclusion principle.* The exclusion principle applied in the game called *musical chairs*. (Source: wikiHow)

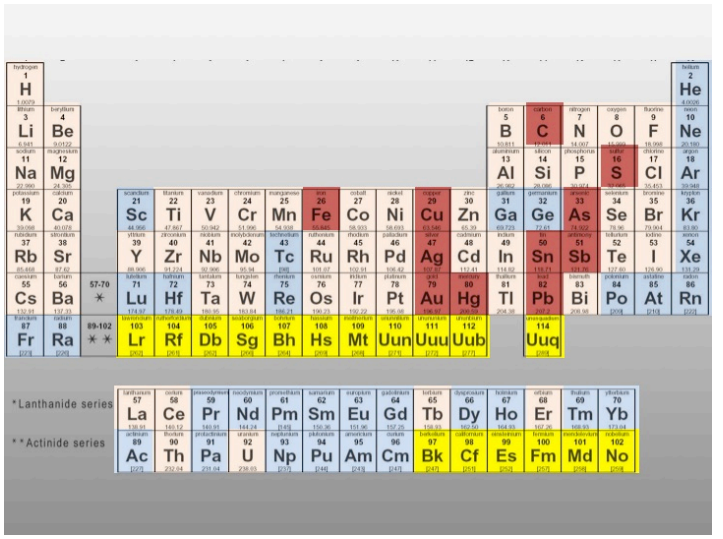


Figure I.4.13: *The struggle to unravel structure.* Mendeleev's periodic table of chemical elements in a historical perspective. The dark red color indicates the elements which were known already in antiquity. Adding the light pink entries you arrive at Mendeleev's table. Including the blue colored elements brings us up to 1945 (Seaborg's table) and the yellow elements were discovered after that. Many entries are thus post-Mendelevian. (Source: Sandbi - Wikimedia Commons)

ties of the microscopic constituents, which are specifically quantum mechanical and have a strong bearing on interactions between the particles, and therefore also on their collective behavior. Consequently there are many macroscopic phenomena which are not obviously quantum, but nevertheless can only be understood if one takes the underlying quantum physics into account.

Going from the microscopic to the macroscopic domain does not necessarily erase all quantum traces. A striking example is the property of spin and the *exclusion principle* of Pauli that – as we mentioned – decrees on the quantum level that particles with half-integral spin cannot occupy the same quantum state. We will have much more to say about this in Chapter II.5 of Volume II , but for the moment we will state the basic facts about it. Whereas the photon is a boson, the electron is a fermion and so are the proton and neutron. So, fermions don't like each other, they like to claim territory and chase away intruders, and they not only try but *have* to avoid each other. In spite of having no genes they certainly come across as rather selfish! Fermions are permanently involved in playing a kind of musical chairs (see Figure I.4.12). For bosons the behavior is the opposite, if the system is at very low temperature and there is no energy to excite the bosons, they love to join each other, and all settle in the same ground state. They will form what is called a *condensate*, a *Bose-condensate* to be specific. These are macroscopically coherent collective quantum states which may exhibit spectacular properties. This form of quantum coherence manifests itself in for example a laser beam, but also in phenomena like superfluidity and superconductivity. We will return to these properties in Chapter III.3 of Volume III.

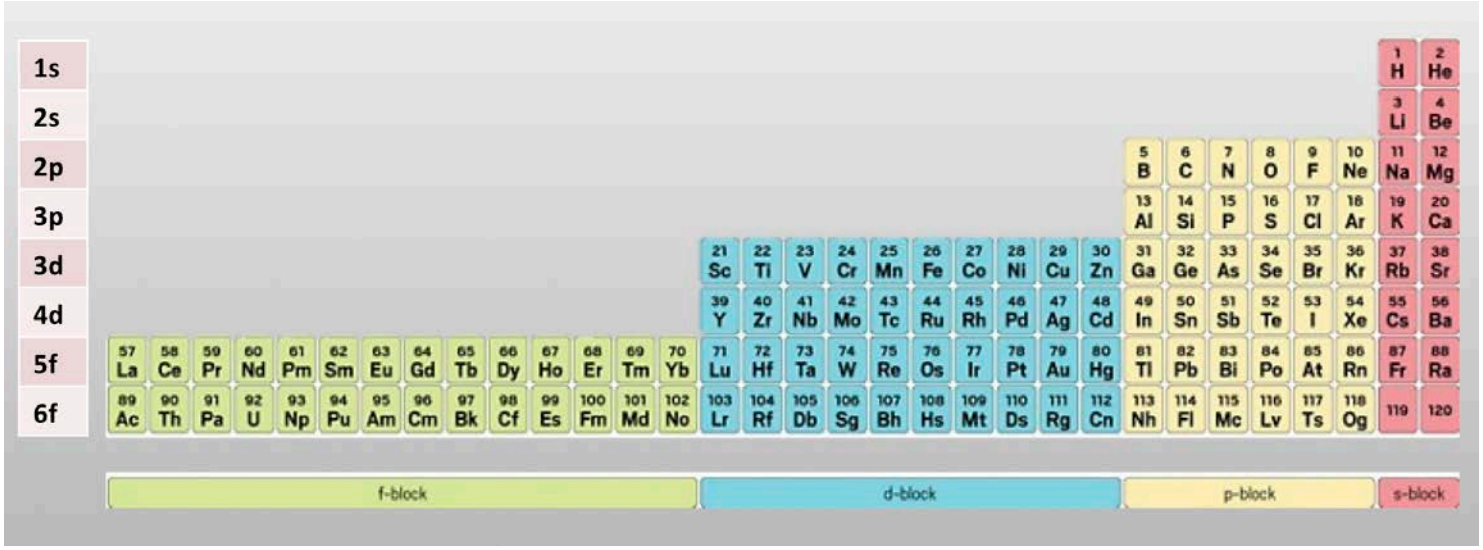


Figure I.4.14: *The left step Janet periodic table.* This is a logical representation of the periodic table in direct correspondence with the quantum classification of atomic states, as depicted in Figure I.4.9. Starting from the right each block corresponds to an increasing integer value of the angular momentum as is indicated in the left column, where letters are used with $\mathrm{l} = 0 \leftrightarrow \mathrm{s}, 1 \leftrightarrow \mathrm{p}, 2 \leftrightarrow \mathrm{d}, 3 \leftrightarrow \mathrm{f}$. And each block contains $2(2\mathrm{l}+1)$ states, i.e. $2, 6, 10, 14, \ldots$. In comparison with the standard Mendeleev representation of Figure I.4.2(a), one sees that the extra rows added (of the Lanthanides and the Actinides) at the bottom of the Mendeleev table get a natural place as $\mathrm{l} = 3$ blocks in the Janet table.

Atoms: the building blocks of chemistry

The connection between Mendeleev's periodic table of the chemical elements depicted in Figure I.4.13, and the systematics of atomic states of Figure I.4.9 is not immediate. This is so because the Mendeleev table was conceived prior to quantum mechanics. This begs for alternative representations of the periodic table in which the underlying quantum structure is manifest.

The rich structure of the periodic table of atomic elements underlying all the structural diversity of chemistry is a direct consequence of the fermionic nature of the electron. Because in an atom with more than a single electron, not all the electrons can sit in the lowest possible state, they have to successively fill the higher energy states of for example Figure I.4.6. And as the chemical properties of the elements are mostly determined by the outer electrons, they will be different because the charge distributions associated with the states of the outer electrons may have quite different shapes, as we saw in Figure I.4.9.

With our knowledge of quantum theory we might prefer to draw the periodic table differently, for example following Janet as we did in Figure I.4.14. In that non-standard visualization there is a direct correspondence with the way the atomic quantum states are labeled as we depicted in Figure I.4.9. The quantum states at a given energy $\mathrm{E_n}$ are organized into *shells* labeled by the angular momentum quantum number l that runs from 0 to $\mathrm{n} - 1$. Quantum theory tells us how things work on the microscopic scale but as a consequence thereof it leaves indelible marks on much larger scales in fields like chemistry and material science. We will have to say a lot more on chemistry and condensed matter in Volume III.

Nuclear structure

Nuclei consist of a certain number of protons and neutrons that are kept together by the strong nuclear force. Nuclei that occur in nature are relatively stable for the excellent reason that they wouldn't be there otherwise, but not all nuclei we find in nature are stable. There are many metastable isotopes *that can decay in a variety of ways, either by the emission of protons, neutrons, α particles, or by $\beta^{\pm}$ or γ radiation. Many of these occur spontaneously in nature and have important applications, for example in the context of carbon dating. Short-lived β^{+} radiators are for example used as radioactive tracers for PET scanning purposes.*

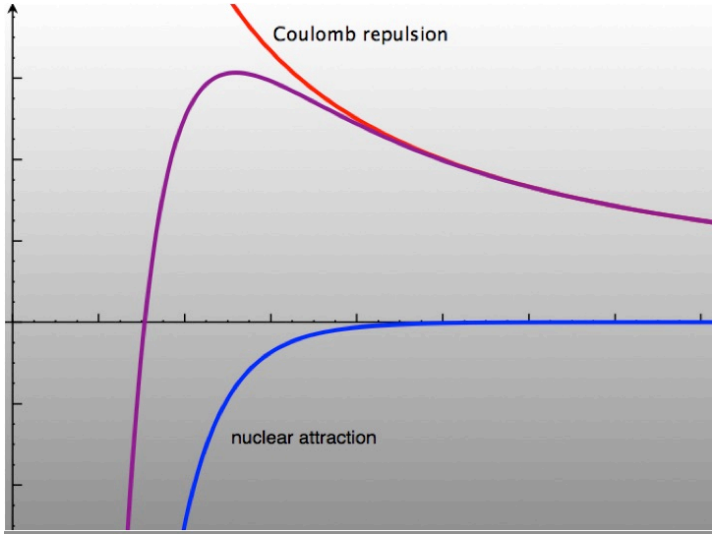


Figure I.4.15: *The nuclear potential between protons as a function of their distance.* The potential is given by the purple curve, which is the sum of a long-range electromagnetic repulsion (in red) and a short-range attractive part due to the strong nuclear force (in blue). Once the particles get close enough they are strongly bound.

An atom consists of a positively charged massive nucleus in the core and a number of electrons 'orbiting' around it, making the overall charge of the atom zero. The natural next step in the quest for fundamental building blocks was to proceed to the structure of the nucleus itself. As always in science, if one observes regularities in structure, one tries to figure out an underlying mechanism that explains those regularities. Here it was not different. The question was open ended in the early days of quantum theory, and it might have happened that one entered a realm where even quantum theory would fail. How exciting! But alas, that didn't happen, physics in the nuclear domain appeared to fully obey the quantum laws. The mechanism underlying nuclear binding is similar to that of the atom in some aspects, but different in others.

Nucleons: Protons and neutrons. Nuclear fission experiments demonstrated that nuclei are composed of particles called *nucleons*, of two types, the *proton* or the *neutron*. From Table B.4 at the end of the book about the discoveries of fundamental particles, we learn that the neutron was discovered by James Chadwick as late as 1932, for which he received the Physics Nobel prize in 1935. but remarkably we also learn nothing about the discovery of the proton as such. That discovery was implicitly made with the discovery of the atomic nucleus by Rutherford in 1911, where the proton is defined as the nucleus of the simplest atom, hydrogen. Rutherford, the great physicist and chemist from New Zealand who spent most of his active research years in Canada and Britain, is often called the 'father of nuclear physics.' He was awarded the Nobel prize for Chemistry in 1908, 'for his investigations into the disintegration of the elements, and the chemistry of radioactive substances.' The neutron was discovered relatively late, presumably because it is unstable as a free particle: it decays under emission of an electron (and an invisible (anti-)neutrino) into a proton! This process was at the root of all radioactive β decay processes of nuclei, discovered by Henri Becquerel as early as 1896 and dramatically expanded by Marie and Pierre Curie. If nuclei are made of protons and neutrons, the first question that comes to mind is: how can positively charged pro-

tons stick together so closely in a tiny nucleus if they all carry the same positive charge? Equal charges repel and repel more strongly if they get closer to each other, because the Coulomb force is inversely proportional to their squared distance. So, how come nuclei don't fly apart? What keeps them together?

A looming crisis leading to a considerable number of gifted Desperado's in search of new physics! A simple but bold idea would be to bluntly postulate a new *strong 'nuclear' force* that would be stronger than the electromagnetic force so that it could overcome the electromagnetic repulsion and cause a net attraction. If we in addition assume that this strong nuclear force works equally strongly on protons and neutrons, this could in principle explain the nuclear binding. And indeed, that is the way it worked out!

The picture looks like Figure I.4.15, where we have plotted the interaction energy of two protons as a function of their distance. It is important to note that there are two contributions, one from the electromagnetic repulsion (the red curve), which is long range and typically falls inversely proportional to the distance, and one from the attractive nuclear force (the blue curve), which is strong but acts only over a short range. These two contributions add up to the interaction energy corresponding to the purple curve where one sees that the repulsion dominates for large distances. Compare these curves for the nuclear binding energy with those we gave for the atomic binding in Figure I.4.5, where the ingredients are similar but work out very differently; in the atomic case the attraction dominates the long distance behavior.

Of course also the instability of neutrons had to be included into this picture as well, and that involves postulating yet another force, the so-called *weak nuclear force*, which will be discussed on page 196.

Isotopes and nuclear decay modes

Isotopes are nuclei that differ from their standard stable composition by having more or less neutrons. This means that these are metastable under various forms of emission. Some are short-lived, and some are long-lived. Nuclear isotopes have important applications.

Isotopes. Nuclei are characterized by two labels, one is the *atomic number* (basically the nuclear charge in units of the elementary charge e) and the other is the *mass number*. These labels can be easily converted into the number of protons, n_p, and the number of neutrons, n_n, in the nucleus, as follows

$$\text{atomic number} = n_p \tag{I.4.9}$$

$$\text{mass number} = n_p + n_n \tag{I.4.10}$$

The basic question was to understand the stability of the well-known atomic nuclei corresponding to the chemical elements. It turned out to be a matter of striking balances. For a chemical element the atomic number in the periodic table is clearly identified with the number of protons in the nucleus. In principle one would expect that, given the electric charge ($\sim n_p$), there could be different numbers of neutrons and therefore one could expect different atomic weights for a given element. This is indeed the case and we speak of different *isotopes* of the element, where the atomic number is the same but the mass number differs. As their charge configuration would be the same, their chemistry would also be, because that is basically governed by the electronic states around the nucleus. Well-known examples of isotopes are *deuterium* and *tritium*, the heavy forms of hydrogen. In addition to the proton, they have one and two neutrons respectively, and are therefore often denoted as $^2\mathrm{H}$ and $^3\mathrm{H}$ as to distinguish them from ordinary hydrogen, $\mathrm{H} = {}^1\mathrm{H}$.

Another important isotope is the carbon isotope ^{14}C, to

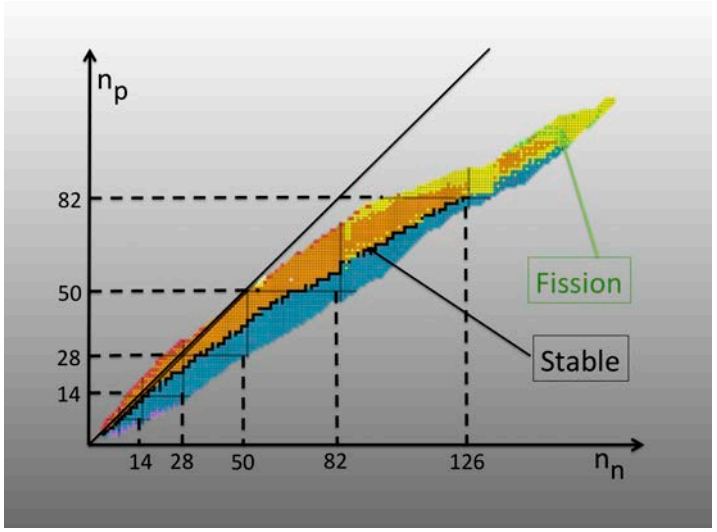


Figure I.4.16: *Stable and unstable isotopes.* The array of nucleotides or nuclear isotopes with on the horizontal axis the number of neutrons and on the vertical one the number of protons. The narrow black band in the middle of the colored region marks the stable nuclei. The other colors refer to nuclear decay types explained in the next figure.

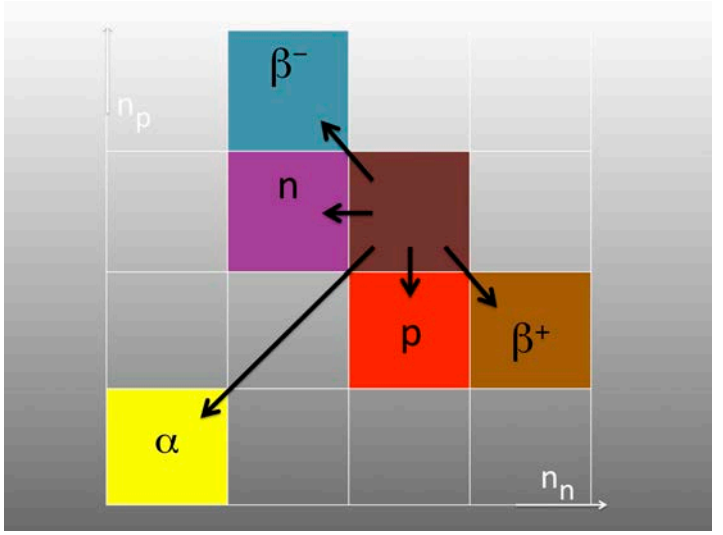


Figure I.4.17: *Nuclear decay modes.* The basic decay modes of nuclei correspond to moves in the diagram: β^- decay corresponds to the emission of an electron, β^+ to the emission of a positron. α-radiation corresponds to the emission of a ^4He nucleus consisting of two protons and two neutrons.

be distinguished from the stable isotope ^{12}C. The former occurs naturally but is unstable, due to the decay-process into nitrogen-14,

$$^{14}C \Rightarrow {}^{14}N + e^- + \bar{\nu}\,, \qquad \text{(I.4.11)}$$

where it emits an electron and an anti-neutrino. This decay is very slow with a half-life $\tau_{1/2}$ of 5730 years. It is this slow decay that is put to use in *carbon-14 dating* methods to determine the age of sediments, fossils and antique art objects. How nice, a nuclear instability that renders an important service to society, as it helps to unambiguously separate real from fake when it comes to providing quantitative, archeological, historical and anthropological evidence about the age of objects.

In Figure I.4.16 we display the array of isotopes, with the number of neutrons on the horizontal axis and the number of protons (i.e. atomic number) on the vertical one. The stable nuclei form the black curve through the center of the colored band, below and above is a band of unstable nuclei that may or may not occur in nature. Note that the line of stable elements is below the $n_p = n_n$ line, which indicates that ever more neutrons are needed to stabilize the nucleus with increasing charge. The line of stable elements ends, indicating that beyond a certain atomic number all isotopes become unstable (around $n_p = 82$).

Nuclear decay modes. At any point in the chart of isotopes there are a number of conceivable instabilities corresponding to moving to neighboring spots as indicated in Figure I.4.17. The nearest neighbors, found by moving, down or sideways in the chart, correspond to adding or getting rid of a single neutron or proton. But we may also think of other so-called *transmutation modes*; for example the nucleus may emit α-radiation, which just means that it emits a (stable) ^4He nucleus consisting of two protons and two neutrons. In our diagram this implies that the nucleus

moves two steps to the left and two steps down. Another possibility is that a neutron in the nucleus decays by β^- decay into a proton and emits an electron (and an anti-neutrino) in which case we move one step up and one to the left. This is because the net charge increases by one unit, meaning that the nucleus would move one step up in our chart, but at the same time it moves one step to the left as the number of neutrons is decreased by one.

For each isotope the dominant decay mode is color coded in the chart of Figure I.4.16, and as expected the dominant decay tends to move the isotope to the black line of stable elements.

The chart shows that away from the stable nuclei marked as black, we have a rather broad band of metastable nucleotides or isotopes, but that band is bounded. On the very right of the table we get into a region where the would-be elements have no stable isotopes at all. These are compounds that do not occur in nature. But that didn't keep physicists like Glenn Seaborg at Berkeley from cooking them up in the lab. And as you see the nuclear physicists have filled out the periodic table up to an atomic number of about 120 by now. The new elements carry legendary names like *Einsteinium, Curium, Bohrium,* and so on. An ironic footnote is, that, while named after scientists whose names may well live forever, the corresponding elements themselves are only extremely short-lived.

Half-lives count. We mentioned already in passing the quantessential notion of a half-life or a decay time usually denoted as τ and it may deserve some explanation. If we take for example a number of N of the of metastable ^{14}C nuclei which have a certain probability to decay, then the number of nuclei that will decay will be proportional to N. This statement can easily be translated in an equation for the decay rate per unit time dN/dt:

$$\frac{dN}{dt} = -N/\tau\,. \qquad \text{(I.4.12)}$$

The solution (see also the *Math Excursion* on page 612 of Volume III) can be written as

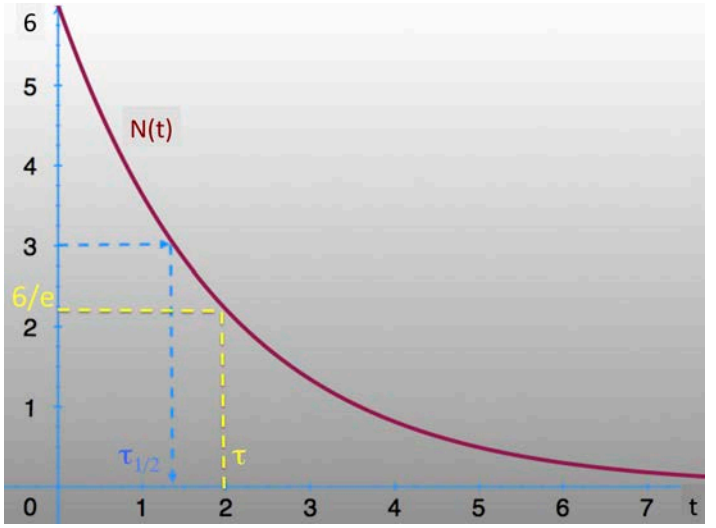


Figure I.4.18: *Half-life versus decay time.* In the case of exponential decay, the half-life $\tau_{1/2}$ is the time needed for half of the initial number N_0 particles to have decayed. After a decay time τ, (which chose equal 2) only N_0/e are left. In the figure we have chosen a scale $N_0 = 6 \times 10^{large}$.

$$N(t) = N(0)e^{-t/\tau}\,, \qquad \text{(I.4.13)}$$

where $N(t)$ is the number of ^{14}C nuclei at time t. You see that the decay is exponential, and the rate equals $1/\tau$. The reader may be more familiar with the notion of a *half-life* $\tau_{1/2}$, the relation is simply $\tau_{1/2} = \tau \ln 2$. This makes the decay go like $2^{-t/\tau_{1/2}}$. so that after time $t = \tau_{1/2}$ only half the number of nuclei are left. In Figure I.4.18 this relation is visualized. What is remarkable about nuclear decays is that their half-lifes can be immense, even as big as the lifetime of the universe! How can a microscopic mechanism with very short characteristic timescales like inside the nucleus produce such incredibly slow processes. Thinking quantum mechanically you would expect a ground state of a certain energy E_0 to typically oscillate with a frequency of order $\nu = h/E_0$, and for a nucleus $E_0 = 1\ keV$ which yields a frequency of $10^{17}\ Hz$ or an oscillation time of order $10^{-18}\ s$. This value has to be contrasted with the decay

time of order 10^{11} s for carbon for example. That is a factor of about 10^{29}!

Imagine: you are an electron and want to get out and you only succeed after banging on the door 10^{29} times! Indeed it is exponentially hard to escape because there is a high potential barrier that wants to keep you inside, the decay is exponentially suppressed, because it proceeds via a process called *quantum tunneling*, a fully quantessential mechanism that has no classical analog at all. Classically the electron would have to climb over the mountain, but it has not enough energy to do that, and there is no escape possible. But quantum mechanically it is more subtle because there is a 'certain uncertainty' in the energy of the particle thanks to Heisenberg. This means that a version of the uncertainty relations applies, reading $\Delta E \cdot \Delta t \geq \hbar/2$. You may loosely paraphrase it by saying that the particle can 'borrow' energy for a brief period of time. It's like magic, if you do the trick fast enough nobody will notice and miracles are possible! Anyhow this means that there is a small probability that the electron will have sufficient energy to get away. That probability is exponentially small though, and depends on the height and width of the barrier. And that explains the enormous factor 10^{-29}. We will say more about quantum tunneling in Part II of the book.

Positron-emission tomography (PET)

Positron-emission tomography is a medical imaging technique for diagnostic purposes. In particular to learn about the functioning of organs. It makes use of specific radioactive isotopes that are injected into the patient. The scanner then traces how the radioactive component is transported through the body.

With the use of isotopes in the medical arena one certainly wants to reduce the exposure of patients to potentially harmful radiation and therefore the isotopes needed for this purpose are typically short-lived positron (β^+) emitters. So here it is anti-matter that matters! If a positron is emitted, it will run into an electron in the detector, and together they will annihilated into a pair of high-energy photons that move out back-to-back. These photons get detected and from their momenta one can reconstruct where the positron was located.

The suitable radio isotopes are thus to be found in the orange region under the black line of stable nuclei in Figure I.4.16. Typical isotopes with short half-lives are carbon-11 ($\tau_{1/2} \sim 20$ min), nitrogen-13 ($\tau_{1/2} \sim 10$ min), oxygen-15 ($\tau_{1/2} \sim 2$ min), or fluorine-18 ($\tau_{1/2} \sim 110$ min). These so-called *tracers* are added to compounds the body uses normally, such as sugars, water and sometimes just the air we breathe (oxygen-15).

Transmutation: Fission and fusion

> Nuclei aren't good or bad, it's what people do with them we have to worry about.

We discuss the basics of nuclear fission and fusion processes, emphasizing their peaceful applications. This includes the large global initiative, ITER, to construct a working net energy producing fusion reactor.

In Figure I.4.19 we show the binding energy per nucleon as a function of atomic mass number. The natural tendency is to minimize the energy: the system will minimize its total binding energy assuming there are no unsurmountable energy barriers that block access to that minimal energy configuration. The graph clearly shows the remarkable and important fact that elements of low mass number tend to lower their binding energy through *fusion* into heavier nuclei, whereas on the other side we see that at high mass number, nuclei can lower their binding energy

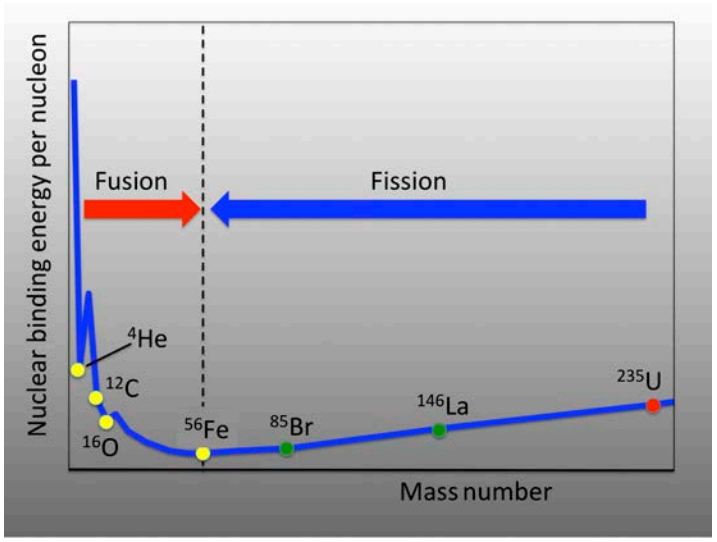


Figure I.4.19: *Fusion and fission.* The binding energy per nucleon inside a nucleus as a function of its mass number. Favored processes are those where this binding energy per nucleon decreases. The light elements tend to fuse, while the heavy ones tend to break up.

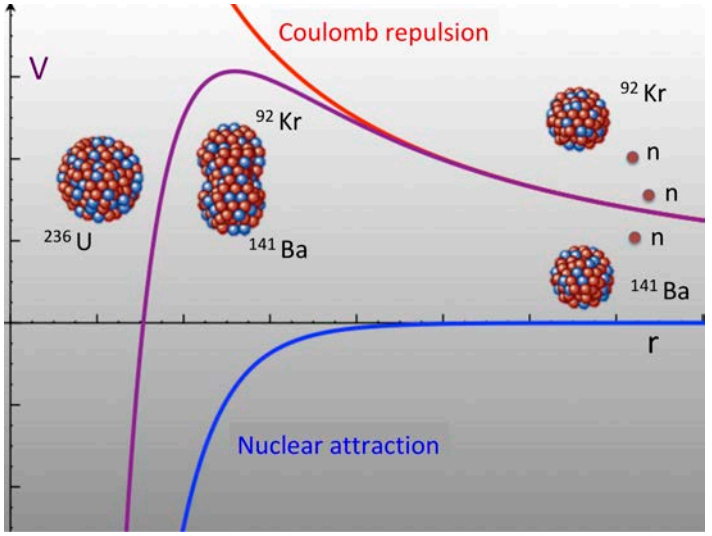


Figure I.4.20: *Fission of uranium,* By absorbing a neutron the ^{235}U isotope changes to the unstable uranium isotope ^{236}U that splits into a ^{141}Ba and a ^{92}Kr nucleus plus three neutrons.

by decaying or *fission* into lighter nuclei. Note also that interestingly the elements ^{4}He, ^{12}C and ^{16}O are relatively stable. In the following subsections we will focus first on fission and then on fusion.

Fission.

> The fundamental point in fabricating a chain reacting machine is of course to see to it that each fission produces a certain number of neutrons and some of these neutrons will again produce fission.
>
> *Enrico Fermi*

The heavy elements on the right of Figure I.4.19 with a high binding energy per nucleon are typically unstable with respect to *decay* or *fission* processes. In these processes the total mass number $(n_p + n_n)$ has to be conserved. We start with fission because it was easier to achieve than fusion – not only in reactors, but also in rather singularly dramatic experiments – like the making of nuclear bombs. In applications, whether it is in the deplorable nuclear weapon industry, or in fission reactors, or in hospitals, one always needs nuclei that are 'fissible'. 'Good fissibility' means that their fission after absorbing a neutron will also produce, apart from the heavy fission products, additional neutrons that can then destabilize neighboring nuclei. This way one can start a chain reaction. And clearly if that is not extremely well-controlled it will turn into an exponentially growing decay process, a meltdown or nuclear explosion, depending on the circumstances. History bears witness to quite a few of such cataclysmic events, and nuclear safety and disarmament should remain a primary concern for all of us. We have to find a responsible balance between profitability and safety and the price is high if we don't get it right.

In Figure I.4.20 we have illustrated the fission of a uranium-235 nucleus after the absorption of a neutron into the nuclei of barium-141 and krypton-91 plus three neutrons. The nuclear process is given by the following reaction equa-

tion:[3]

$$ {}^{1}_{0}n + {}^{235}_{92}U \rightarrow {}^{141}_{55}Ba + {}^{92}_{36}Kr + 3\,{}^{1}_{0}n\,. \qquad \text{(I.4.14)} $$

It is clear that the emitted neutrons can ignite other uranium nuclei and this will keep the process going, provided that there is a sufficient concentration of uranium-235.

Natural uranium is found in ore deposits in many places around the world. It is predominantly a mixture of the two isotopes uranium 238 (99.27%) and uranium 235 (0.72%), and therefore to make nuclear fuel that can be used in reactors, one has to increase the 235 fraction by an 'enriching' process, for example by using centrifuges to get rid of the heavier 238 isotope. In a fission reactor the process is moderated by neutron absorbing materials such as graphite, water or heavy water (where the hydrogen is replaced by deuterium). The uranium-235 itself has a natural half-lifetime of 703,800,000 years, so no wonder there is still a lot left from the original amount stocked in the Earth crust. It naturally decays by emitting an α particle, producing a thorium-231 which in turn then decays rapidly in protactinium-231 and so on. It winds up in a long chain of successive reactions of which some are fast and others slow, with half-lives of thousands of years. The reaction chain of uranium-235 ends with the element lead-207 $({}^{207}_{82}Pb)$, which is stable. However, if we get the uranium-235 to absorb a neutron, it turns into a uranium-236, and that is unstable so it breaks up in krypton and Barium plus three neutrons, and that can keep a chain reaction going.

Fusion. Going back to the binding energy curve of Figure I.4.19 we now turn to the left side, where we see that energy can be gained if we manage to *fuse* light nuclei (like hydrogen) into a stable nucleus with higher atomic number (like helium-4). This is not so simple because one has to 'overcome' the electromagnetic Coulomb repulsion between for example two protons. Now in an accelerator this certainly could be done but to do this on a larger scale one has to achieve physical conditions which are quite extreme. So, to get fusion going has turned out to be very, very difficult. In spite of numerous experts who have been raising expectations, the timescales for achieving fusion have been repeatedly extended by decades. To go from 'scientific feasibility' to 'successful technology' sometimes takes a long time and may be hard to estimate. This leads to the familiar situation where either the optimists or the pessimists are ridiculed!

The Lawson criterion. How extreme the conditions are that have to be met in order to get fusion to work can be

Chrysopoeia: transmutation into gold?

There was a lot more to magic, as Harry quickly found out, than waving your wand and saying a few funny words.
J.K. Rowling, Harry Potter and the Philosopher's Stone.

Making gold is the alchemist's dream! In alchemy, the term *chrysopoeia* means transmutation into gold. It comes from the Greek words χρμσοσ, khrusos, meaning 'gold,' and ποιειν, poiëin, meaning 'to make.' The term refers to the creation of the *stone of wisdom* or the *philosopher's stone*. In the early days of alchemy, in Egypt and Greece there was a serious quest for the stone, as it would allow you to turn any metal into gold. It apparently led to a kind of primordial gold rush. For example, Zosima's *formula of the crab*, supposedly constituted a kind of recipe to brew gold out of copper and zinc. If only copper, zinc and a *Bunsen* burner would do! This 'recipe' would instantly turn any 'nitwit' into a billionaire, for as long as they

[3] We use the notation ${}^{A}_{Z}X$ with X= chemical element, A= mass number and Z= atomic number.

managed to keep it secret of course! In Egyptian antiquity there must have been loads of books on alchemy, and – from a historical point of view – unfortunately, almost all of them have been lost. It was an 'executive order' by the Roman Emperor Diocletian in AD 296, which decreed that all alchemy books on making gold had to be burned. Anyway, we all know that the true heirs of alchemy are of course our friends the stockbrokers. Or should I say the Silicon Valley tech-billionaires who turn doom scrolling addictions into gold!

Now you ask, can nuclear physics revive the old gold-plated dream in a more mundane way? The answer is a clear 'yes!' Gold was first synthesized from mercury by neutron bombardment in 1941, but the isotopes of gold produced were all radioactive, so the gold produced had an expiration date, and that is precisely what you don't want. You don't want a fragile 'bread and butter' like commodity to be your gold standard. Actually there is only one stable gold isotope, $^{197}\mathrm{Au}$, so to produce desirable gold, nuclear reactions must create this isotope.

It can be done, but unfortunately it is way more expensive than just buying gold. Gold can actually be manufactured in a nuclear reactor by irradiation of mercury with neutrons. For this to work you need the mercury isotope $^{196}\mathrm{Hg}$, which occurs with a frequency of 0.15% in natural mercury. That isotope can be converted into gold, by first absorbing a neutron and then through electron capture decaying into $^{197}\mathrm{Au}$ with some slow neutrons. I think we can be sure that, all those painfully negotiated and maintained nuclear nonproliferation agreements are not made out of fear that bad people might embark on breeding a nuclear goose producing golden eggs *ad infinitum*. □

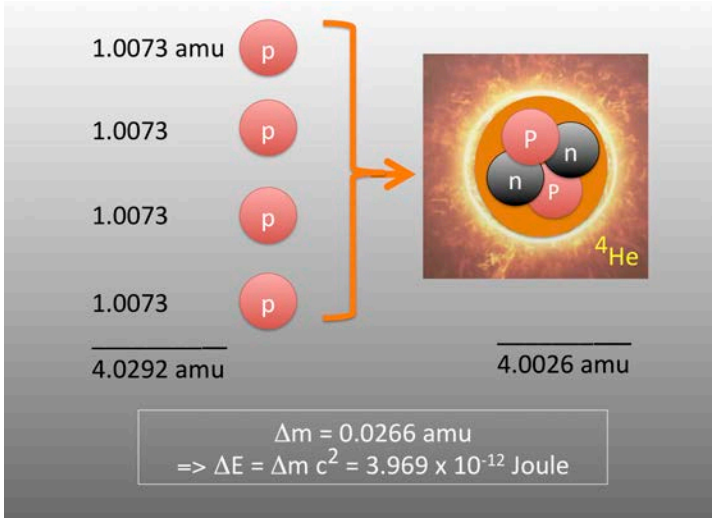


Figure I.4.21: *Energy gain by fusion.* The net energy gain from a fusion of four protons into a $^4\mathrm{He}$ nucleus, as it happens in the Sun. One *atomic mass unit* or amu corresponds to $931.5\ \mathrm{MeV/c^2} = 1.661 \times 10^{-27}\ \mathrm{kg} \leftrightarrow 1.492_1 0^{-}10\ \mathrm{joule}$.

expressed by the so-called Lawson criterion. John Lawson, a young engineer working on nuclear fusion, decided in 1955 to work out exactly how hard it is to achieve fusion. Although his colleagues were quite optimistic about their prospects, he wanted to prove it to himself. He found that the conditions for fusion power relied on three vital factors. By calculating the requirements for more energy to be created in the plasma than is put in, he came up with a dependence on three quantities: temperature (T), density (n) and confinement time (τ). He derived a lower bound on the triple product, $\mathrm{L} \equiv \mathrm{n}\,\tau\,\mathrm{T}$ which would depend on the type of process and the type of machine. For the deuterium-tritium fusion one typically needs $\mathrm{L} \geq 10^{21}\ \mathrm{keV\ s/m^3}$ and that is what the international fusion project ITER in France is expected to achieve. The technological promise of a fusion reactor based on the Tokamak concept, where an extremely hot nuclear plasma is confined to a toroidal reaction chamber by very strong magnetic fields, has been clearly established. So far no stable, net energy producing fusion device has been constructed,

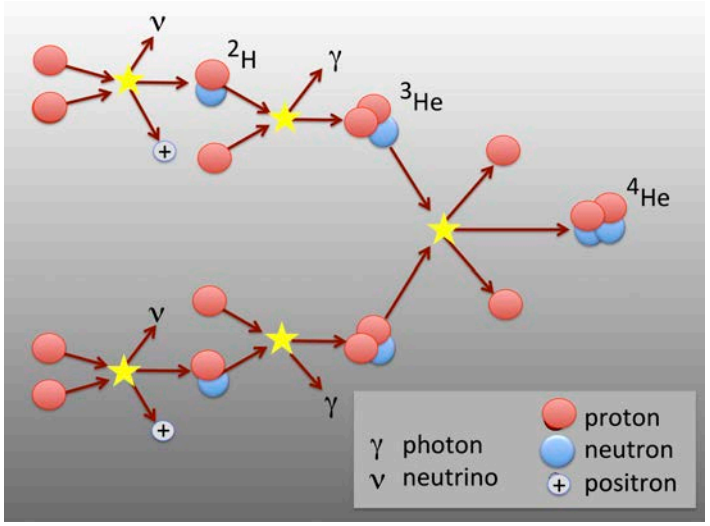


Figure I.4.22: *Fusion in the Sun.* This is the chain of nuclear fusion processes that takes place in the core of the Sun. It is a three step process, leading from protons via deuterium and helium-3 to the stable helium-4 nuclei. The net result is that four protons are converted into a single helium-4 nucleus.

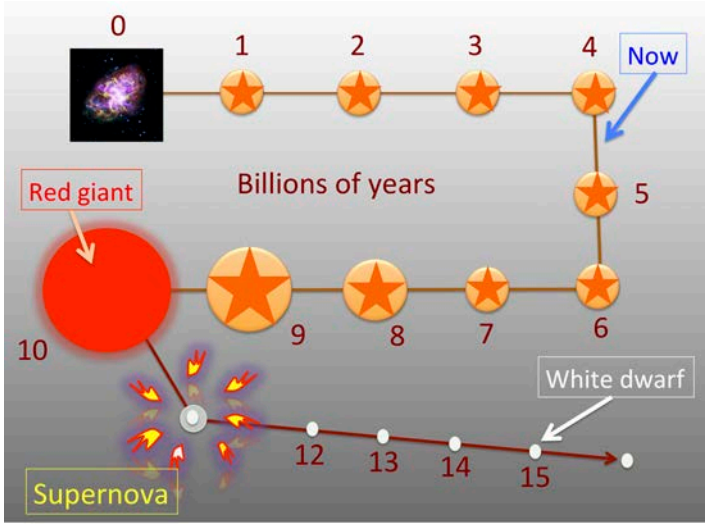


Figure I.4.23: *The life cycle of the Sun.* An average star like the Sun has sufficient hydrogen to burn by fusion so as to keep shining for about 10 billion years. It will then form a red giant after which the core collapses to a white dwarf about the size of the Earth.

but we will discuss the ITER project shortly. It is a sobering thought that it is not us who invented fusion, of course nature did. And we have learned a lot from studying and understanding the energy production in the Sun which basically is a gigantic nuclear fusion reactor.

Let the Sun shine.

> ... No more falsehoods or derisions
> Golden living dreams of visions
> Mystic crystal revelation
> And the mind's true liberation ...
> Let the sun shine, let the sun shine in!
>
> *The fifth dimension* in the musical *Hair (1967)*

The extreme pressure caused by the gravitational force in the core of stars turns them into extreme pressure cookers, allowing for all kinds of fusion processes to take place. Every second, our Sun turns 600 million tons of hydrogen into helium, releasing an enormous amount of energy. Achieving fusion on Earth has required a different approach since we lack a natural pressure cooker to achieve the densities and temperatures needed. The temperature at the Sun's surface is 6,000 degrees, while at its core it is 15 million degrees. Temperature combines with density in the Sun's core to create the conditions necessary for the fusion reaction to occur. The gravitational forces of our stars cannot be recreated here on Earth, and much higher temperatures are necessary in the laboratory to compensate.

The basic process of burning hydrogen to produce helium through the chain of fusion processes is depicted in Figure I.4.22. The hydrogen nuclei are just protons, so, in the first step we make deuterium under emission of a neutrino and a positron. The second step is to have the deuterium and a proton fuse into helium-3 under emission of a photon. Finally two helium nuclei can fuse into the stable

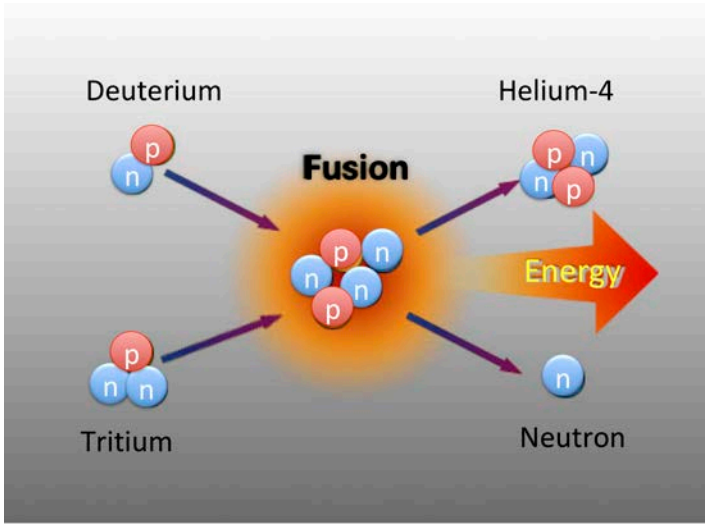


Figure I.4.24: *The basic ITER process.* The basic fusion process: ${}^2\mathrm{H} + {}^3\mathrm{H} \rightarrow {}^4\mathrm{He} + \mathrm{n}$ is delivering the energy in the ITER fusion reactor. The difference in mass between the total incoming and out-coming nuclei is converted into energy according to Einstein's formula $\mathrm{E} = \mathrm{mc}^2$.

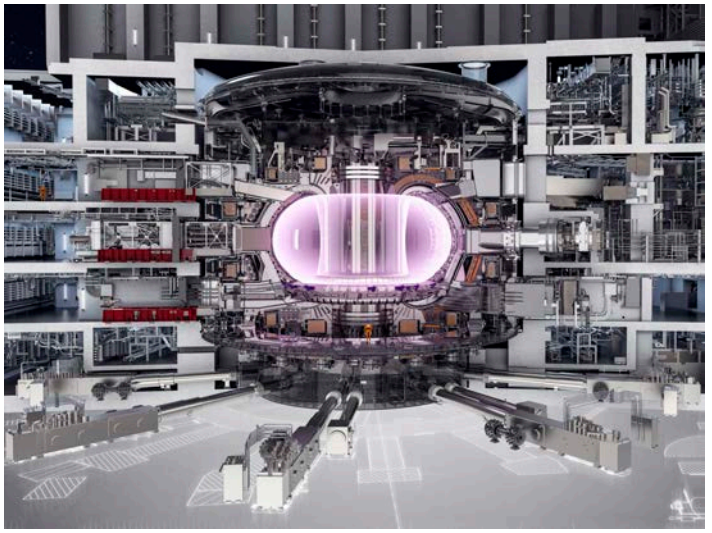


Figure I.4.25: *ITER.* The international fusion reactor located in France. The reaction chamber contains the plasma which is enclosed in a toroidal magnetic field configuration, where it is heated up to temperatures of a few hundred million degrees so that fusion can take place. (Source: ITER)

helium-4 nucleus under emission of two protons. The net energy delivered in such a process is what is calculated in Figure I.4.21: it amounts to about 4.0×10^{-12} joules per helium-4 nucleus produced. In other words, burning 1 kg of fuel this way would produce about 2.3×10^7 MWh. This is comparable to what a 100 MW energy plant produces in 26 years!

Having analyzed the energy production of the Sun, we have also answered the question whether the Sun will keep shining forever. The answer is a firm 'no', because the Sun will simply run out of fuel at some point. The long-term perspective for life on Earth looks quite dim. In some 5 billion years the Sun will first blow up to form a *red giant* that will swallow the inner planets (including the Earth). The core will then collapse to a compact stellar object called a *white dwarf* while the outer parts will be blown off in space. The life cycle of the Sun is schematically depicted in Figure I.4.23. So, beware: our days are counted!

ITER: the nuclear fusion reactor

ITER will be the first fusion device to produce net energy and it will be the first fusion device to maintain fusion for long periods of time. Furthermore, it will be the first fusion device to test the integrated technologies, materials, and physics regimes necessary to enable a commercial production of fusion-based electricity.

The ITER project comprises a truly global collaboration, where China, the European Union, India, Japan, Korea, Russia and the United States are now engaged in a 35-year project to build and operate the ITER experimental device. The goal of the program is to produce a net gain of energy and deliver a prototype for the fusion power plant of the future. It has been designed to produce 500 MW of output power for 50 MW of input power – or ten times the amount of energy put in. The current record for released fusion power is 16 MW (held by the European JET facil-

ity located in Culham in the UK). In the ITER Tokamak, where the plasma is confined by strong magnetic fields into a toroidal reaction chamber, temperatures will reach 150 million degrees, that is ten times the temperature at the core of our Sun!

The 180-hectare ITER site in Saint Paul-Les-Durance, in the south of France, has a 42-hectare platform the size of 60 soccer fields. Building began in August 2010. The hope is that the reactor will be completed around 2030.

Field theory: particle species and forces

A major achievement in quantum physics in the second half of the twentieth century is the development of quantum field theory (QFT). It is a general formalism that encompasses both quantum theory and special relativity, and dramatically shifted our perspective on what particles deep down really are. It made us understand the origins of spin*, and of the* exclusion principle *and its related particle statistics properties. These developments culminated in the Standard Model which comprises precise and explicit new theories that describe the strong and weak nuclear interactions, as well as electrodynamics. Quantum theory opened the door to the microcosmos, and quantum field theory appears to correctly describe all processes down to the smallest scales we have been able to probe so far.*

Our quest to understand nature at ever smaller scales, forced us to study elementary processes at ever higher momenta and energies. This is a direct consequence of Heisenbergs uncertainty relations. Making Δx small requires making Δp and thus p and E large. To achieve such extreme energies one had to build big particle accelerators like CERN near Geneva and Fermilab near Chicago. Imagine, the energy consumption of one such machine is comparable to that of a medium-size city!

When the energies become of the same order as the rest masses of the elementary particles involved, one necessarily has to take special relativity into account. In particular, in view of the equivalence of mass and energy, we have to anticipate processes occurring where energy will be converted into mass and the other way around. On the one hand we expect the production of massive particles out of pure energy, and on the other hand the creation of pure radiation energy out of particle anti-particle annihilation. To make further progress in understanding these processes a theoretical framework that is consistent with both quantum mechanical and (special) relativistic principles was needed. The problem was in fact twofold: one was to find the relativistic generalization(s) of the Schrödinger equation, and the other was to develop a formalism for many particle states, where particles could be created and annihilated and converted into pure energy in the form of photons for example. Implementing these two requirements together gave rise to the (relativistic) *quantum field theory* formalism.

Relativistic wave equations. Let me recall that the classical Maxwell theory is already relativistically invariant. In fact, it was electromagnetism that pointed Einstein the way to relativity because it was hidden in there. You could say that the Maxwell equations are relativistic but not really quantum yet. With the Schrödinger equation the problem is the other way around, it is quantum but not relativistic. It is not, because it is based on the Newtonian – therefore non-relativistic – definitions of energy and momentum. We looked at the basics of the Schrödinger equation in a previous section and constructed the Schrödinger wave equation by means of a substitution where we replaced the classical E and p variables with differential operators, as shown in equation (I.4.2).

That exercise showed that the Schrödinger equation is *not* relativistically invariant. It is a wave equation, but quite different from the electromagnetic wave equation (I.1.47), which features the relativistic wave or 'box' operator, we

introduced in equation (I.1.31). Indeed, in the electromagnetic wave equation discussed in Chapter I.1 , space and time are treated on equal footing, and that is not the case for the Schrödinger equation because it has a first-order time derivative, but second-order spatial derivatives.

Naively following the same approach, we could start with the relativistic expression for the particle energy and make the same substitutions:

$$E^2 = \mathbf{p}^2c^2 + m^2c^4 \quad \rightarrow \quad (\hbar^2\Box + m^2c^2)\phi = 0\,. \quad \text{(I.4.15)}$$

Not surprisingly, we now meet again our old friend the relativistic wave operator $\Box$, and in addition a mass term. This seems quite straightforward, and in fact it is. This equation was already written down by Schrödinger himself, who discarded it for reasons that we will point out shortly. The resulting equation is called the *Klein – Gordon (KG) equation*, and after all the dust of field theory settled, it turned out to have a consistent interpretation: it describes a scalar particle, or a particle without spin, such as for example the pion.

No ground state, no physics! On the level of a quantum equation for a single particle, interpreting the Klein – Gordon like the Schrödinger equation, gave rise to a real problem with it. Let me digress a little on what that problem was about. If I tell you that $b = 2$ is true, you may say: 'fine, so be it', then I square that equation and say $b^2 = 4$, and again ask you what is b? Well then, if you, once upon a time, had dutifully executed your homework assignments, you would not answer $b = 2$, but $b = +2$ *or* $b = -2$. By squaring the equation, I have smuggled in an extra negative solution. I managed to somehow double the truth! How shrewd, the logic is impeccable but not always reversible. The quadratic equation is less restrictive.

What this means is that the quadratic relation for the energy (and the corresponding wave operator), in the KG equation also introduces negative energy solutions, after all the solutions are $E = \pm\sqrt{\mathbf{p}^2c^2 + m^2c^4}$. So we do not add just one, but infinitely many negative energy solutions. Well, nothing wrong with that, if we go back to the bound states in the hydrogen atom. we see that also there we had an infinity of negative energy bound states. The significant difference, however, is that the negative energy values obtained from the Klein-Gordon equation are not bounded from below because the magnitude of the momentum is unlimited. In other words, there would be no ground state, and the particle it describes would be unstable. Unfortunately, no ground state means no physics!

People got stuck in the Klein – Gordon theory, because it seemed impossible to interpret satisfactorily. And indeed to do relativistic quantum mechanics correctly, one had to go beyond writing down a wave equation for a single particle. One would have to go to quantum field theory to resolve the apparent inconsistencies with these equations. Nevertheless, the idea of somehow producing a sensible relativistically invariant first-order equation as a kind of 'square root' of the Klein – Gordon equation was on the table, and the hope was that that would resolve the problems of that equation.

The Dirac equation: matter and anti-matter 

> Dirac was the strangest man who ever visited my institute. During one of Dirac's visits I asked him what he was doing. He replied that he was trying to take the square-root of a matrix, and I thought to myself what a strange thing for such a brilliant man to be doing. Not long afterwards the proof sheets of his article on the equation arrived, and I saw he had not even told me that he had been trying to take the square root of the unit matrix!
>
> *Niels Bohr*
>
> (Quoted in Kurt Gottfried, *P.A.M. Dirac and the Discovery of Quantum Mechanics.*)

The remarkable features of the electron, like having an intrinsic angular momentum called spin and being subjected to the mysterious Pauli exclusion principle, all fell into place after Dirac wrote down his beautiful, relativistically invariant, first-order wave equation for the electron. But the biggest surprise was its prediction of the existence of anti-matter.

The relativistic equation for the spin one-half electron was published in 1930, and Paul Adrian Maurice Dirac shared the Nobel prize with Erwin Schrödinger three years later. This equation, for the electron and its anti-particle the positron, and the Maxwell equations describing photons, forms the back-bone of a theory called *Quantum Electrodynamics (QED)*, which constituted the first example of a consistent relativistic quantum field theory. With the completion of this theory shortly after the Second World War, a fully relativistic and quantum mechanical treatment of the electromagnetic interactions of electrons, positrons and photons was achieved.

On taking square roots. To get some appreciation for one of the most beautiful equations of physics, it is illuminating to go back to the Klein – Gordon equation as a starting point. We would like to take the positive root, so to say, of the Klein-Gordon, but that is hard. On the mechanics side on the left of (I.4.15), with the algebraic relation it is easy, you just take the root on both sides and only keep the positive root by choosing[4] $E = +\sqrt{\mathbf{p}^2 + m^2}$. But on the Klein-Gordon side of the story, you would have to take the root out of the $\Box$ operator and that is hard to define, because you have to define what you mean by the square root of a derivative. Strictly speaking you could express it as an infinite series of ever higher powers of the momentum operator but that is not what you want, because that would involve taking 'infinite order derivatives' and that makes even strong people quail! What you really would like to have is an expression *linear* in E, p and m that squares to the Klein-Gordon operator. And that is what Dirac brilliantly achieved by making use of matrices in defining this miraculous 'square root'.

A matrix root: the Weyl equation Let me take one step at a time and first indicate why using matrices dramatically enlarges the space of possibilities for taking a square root.[5] Let us pose 'taking the square root' as a matrix problem. Suppose that instead of the equation $b^2 = 4$, which of course has solutions $b = \pm 2$, I would have considered the matrix equation $B^2 = A$ with A being 4 times the 2×2 unit matrix:

$$A = 4 \cdot \mathbf{1} = \begin{pmatrix} 4 & 0 \\ 0 & 4. \end{pmatrix}$$

If I ask you to solve the equation for B, then you could have come up with 4 independent solutions. If you start with the set $\{X^\mu\}$,

$$X^0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, X^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$
$$X^2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, X^3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \text{(I.4.16)}$$

then 4 independent solutions would be $B^\mu = 2X^\mu$. We may go one step further and check that the much stronger identity holds:

$$(p_0 X^0 + \mathbf{p} \cdot \mathbf{X})(p_0 X^0 - \mathbf{p} \cdot \mathbf{X}) = (p_0^2 - \mathbf{p}^2)\mathbf{1}, \quad \text{(I.4.17)}$$

because of the special properties of the set of matrices $\{X^\mu\}$. Multiplying out the left-hand side you get $4^2 = 16$ terms that are quadratic in both the X-matrices and the momentum components p_μ. Equating the coefficients of the six *different* momentum combinations $p_\mu p_\nu$, one obtains six equations that the matrices have to satisfy. Firstly, we have the condition that the symmetric products or *anti-commutators* of the matrices have to satisfy $\{X^i, X^j\} \equiv$

[4] To make the argument and formulas more transparent we choose natural units where $\hbar = c = 1$ in this subsection.

[5] This calculation uses a little bit of the material out of the *Math Excursion* on complex numbers at the end of Part III. Here it suffices to know that i denotes the 'imaginary unit' and that it by definition squares to minus one: $i^2 = -1$.

$X^iX^j + X^jX^i = 2\delta^{ij}$. Secondly, we have the condition that the antisymmetric products or *commutators* have to satisfy $[X^0, X^j] \equiv X^0X^j - X^jX^0 = 0$. And as you may check, the matrices X^μ do exactly that! This special set of matrices, are called the *Pauli-matrices* or *spin-matrices* that are often denoted as σ_μ. The reason they are called the spin-matrices will become clear shortly.

So, we succeeded in writing the four-momentum squared, as a product of two matrices linear in the momentum, that is without using square roots. But this nice construction can be applied to equations as well. If we have a massless relativistic particle, its momentum satisfies $p_\mu p^\mu = 0$, leading to a massless Klein-Gordon equation for a (spin zero) scalar field of the type $\Box\psi(x^\nu) = 0$. However, with what we just learned one could also introduce a linear first order matrix equation. This is just the so-called Weyl equation, named after the German mathematician, theoretical physicist and philosopher Hermann Weyl, who wrote this relativistic wave equation down in 1929:

$$(iX^\mu\partial_\mu)\,\Psi(x_\nu) = 0\,, \tag{I.4.18}$$

where Ψ is a two-component, so-called *spinor*, on which the matrices work. The wave-like solutions are of the form:

$$\Psi \sim u(p)e^{-ip_\mu\cdot x^\mu}\,, \tag{I.4.19}$$

with $u(p)$ a spinor. Substituting this in the Weyl equation we get an algebraic equation for the two-component spinor $u(p)$:

$$(X\cdot p)\,u(p) = 0 \quad \rightarrow \quad \mathbf{X}\cdot\mathbf{p}\,u(p) = p_0\,u(p)\,. \tag{I.4.20}$$

This is an eigenvalue equation with two independent solutions $u(p) = \eta^\pm(p)$ and eigenvalues $p_0 = E_\pm = \pm|p|$:

$$\mathbf{X}\cdot\mathbf{p}\,\eta^\pm = E_\pm\,\eta^\pm\,. \tag{I.4.21}$$

This positive energy η^+ mode describes a massless particle with spin one-half, with its spin polarized parallel to its momentum. It is a particle with a fixed positive *helicity* which therefore is also called a *right-handed* particle. The negative energy η^--component describes the corresponding *anti-particle* which necessarily has the opposite helicity.
The first factor on the left-hand side of equation (I.4.17), also describes a two-component spinor which can be obtained from the one we just discussed by flipping the sign of the energy p_0, so it will describe a left-handed or negative-helicity particle, and its anti-particle.

The first thing we have to conclude is that the Weyl equation describes a relativistic spin one-half particle. We did however not get rid of the negative energy solutions, but presumably these have to be interpreted as describing an anti-particle. We will return to this picture shortly.

For a long time it was believed that neutrinos would be massless, left-handed particles described by a Weyl equation, but we have in the meantime learned that neutrinos have a small mass after all. They therefore have to be described by a Dirac equation where the two chiralities get coupled through the mass term.

The Dirac matrices and algebra. Dirac managed to do something similar for a massive particle. He started with the quadratic relativistic energy-momentum relation (times the unit matrix), and wrote it as a product of two matrix factors linear in the momentum. To succeed he needed to introduce four 4×4 matrix coefficients γ^μ. Using the standard, very convenient, 'slash' notation $\not{p} \equiv p_\mu\gamma^\mu$ (introduced by Feynman), we may write:

$$(\not{p} + m\mathbf{1})(\not{p} - m\mathbf{1}) = (E^2 - \mathbf{p}^2 - m^2)\mathbf{1}\,. \tag{I.4.22}$$

Again, multiplying out the left-hand side out you get $4^2 = 16$ terms that are quadratic in both the gamma matrices and the momentum components. To satisfy the equation, the diagonal terms require $(\gamma^0)^2 = \mathbf{1}$ and $(\gamma^i)^2 = -\mathbf{1}$, while the six terms with a product of two *different* momentum components should all vanish. The matrix coefficients correspond to the *anti-commutator* of the corresponding

gamma-matrices:

$$\{\gamma^\mu, \gamma^\nu\} \equiv \gamma^\mu \gamma^\nu + \gamma^\nu \gamma^\mu . \qquad \text{(I.4.23)}$$

The upshot is that conditions on the gamma matrices that follow from the requirement that equation (I.4.22) is satisfied are summarized by the equation:

$$\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}\mathbf{1} , \qquad \text{(I.4.24)}$$

where $\eta^{\mu\nu} = diag(1, -1, -1, -1)$ is the relativistic Lorentzian space-time metric we encountered before. Matrices satisfying an algebraic relation like the one above form a so-called *Dirac or Clifford algebra*.

The Dirac equation. We are ready to tackle the four-component Dirac equation which in its most compact and elegant form can be written as:

$$(i\not{\partial} - m\mathbf{1})\psi(x^\mu) = 0 , \qquad \text{(I.4.25)}$$

This first-order system is relativistically invariant, because one can show that the matrices do indeed also transform like a four-vector. It has wave-like solutions multiplied by a four-component *spinor* $u(p)$. The 4×4 γ matrices act on the components of the spinor. For positive energy ($E > 0$) the solutions look like:

$$\psi(x^\mu) \sim u(p) e^{-ip_\mu x^\mu} , \qquad \text{(I.4.26)}$$

substituting this in the Dirac equation yields the algebraic equation for $u(p)$:

$$(\not{p} - m\mathbf{1})u(p) = 0 . \qquad \text{(I.4.27)}$$

The negative energy solutions can be written in a similar way as:

$$\psi(x^\mu) \sim v(p) e^{+ip_\mu x^\mu} \qquad \text{(I.4.28)}$$

and it yields an equation for the spinor $v(p)$:

$$(\not{p} + m\mathbf{1})v(p) = 0 . \qquad \text{(I.4.29)}$$

Comparing these equations we see that the Klein – Gordon equation factorizes into a product of two first-order equations. These two equations are then combined again in the single four-component Dirac equation, which admits positive and negative energy solutions: the former correspond to the electron and the latter to the *hole* (or positron) degrees of freedom respectively.

It is important to remark that the four components of the wavefunction *not* form a four-vector; they form a four-component *spinor* which transforms differently under Lorentz transformations. Another way to say this is, that of the four components, two states correspond to an electron with its two spin states, while the other two would correspond to a positron with its two spin states. But as the gamma-matrices are not diagonal the equation mixes all components. There is a lot of beautiful and important mathematics hidden in the Dirac equation that we will not address here at all. Our goal was to get to know the magnificent equation that provided such a deep understanding of quantessential properties of matter like spin, the exclusion principle and the necessity of anti-matter. ■

The spectrum. Let us first look at the energy spectrum of the free Dirac particle as depicted on the left in Figure I.4.26. The first thing that strikes us in this picture is that the negative energy states have not disappeared. So, again it looks like there is no lowest energy state, and taking the square root of the equation has *not* eliminated the negative energy states in any obvious way. Consequently one would think that this feature would make the model inconsistent and useless. But, no! Dirac brilliantly argued that because his particles necessarily have spin one-half, they would have to satisfy the exclusion principle. But if that is the case, he could decree that all negative energy states would be filled, and there would be no problem. There would be a lowest energy state for the next electron to come in. So Pauli's exclusion principle acts like a *deus ex machina* here.

The second point to observe is that there is an *energy-gap* of $\Delta E = 2mc^2$ between the highest negative energy state

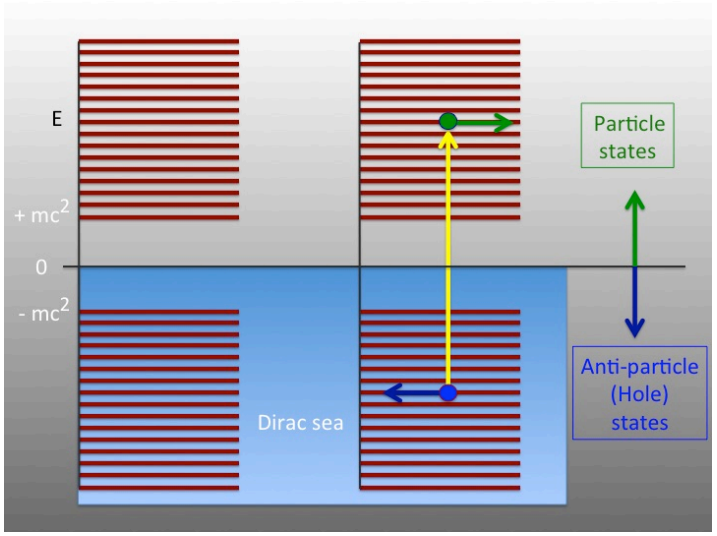


Figure I.4.26: *The spectrum of the Dirac field.* The energy spectrum of the Dirac equation for a spin 1/2 particle of mass m. It has positive and negative energy states. The negative energy states are all filled and form the 'Dirac sea.' A high-energy photon ($\mathrm{E} \geq 2\mathrm{mc}^2$) can excite an electron out of the sea into a positive energy *particle* state, and the hole that stays behind is just an *anti-particle* with opposite charge and opposite momentum.

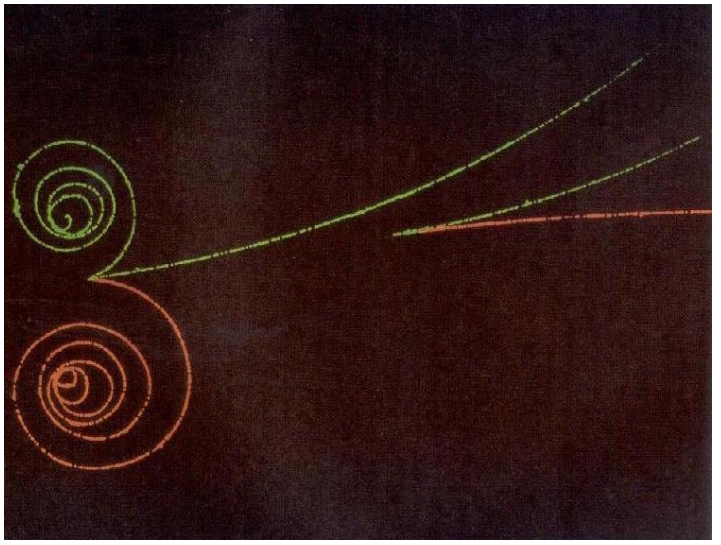


Figure I.4.27: *Pair creation.* This is bubble chamber picture of a high-energy photon which enters from the left, and is not visible because it has no charge). It knocks out an electron thereby also creating a relatively low energy pair. Later the photon produces a second pair with more energy. A strong magnetic field is applied perpendicular to the page, which causes the particle trajectories to curve depending on their charge and energy. A perfect way to split the electron and positron tracks therefore.

and the lowest positive energy state. In the field theory context this means that exciting an electron from a negative energy state to a positive state would cost at least $2\mathrm{mc}^2$, and effectively produce both a *particle and* a '*hole*'. There is no such thing as only creating a particle. The 'hole' is nothing but the *anti-particle* or *positron*, having the same mass and the opposite charge. So from the 'vacuum' state, corresponding to the completely filled 'Dirac sea' of negative energy states, one may create particle anti-particle (electron-positron) or particle-hole pairs. This is indicated on the right-hand side of the Figure I.4.26. A bubble chamber shown in Figure I.4.27 clearly shows the successive creation of two pairs from a high-energy photon. Understanding of the Dirac equation leads therefore inevitably to the prediction and discovery of anti-matter.[6]

Condensed matter. It is quite gratifying to see that 80 years after Dirac wrote his equation down, it is still alive and kicking. This equation is there to stay! Apparently Dirac himself once quipped that the equation was far more intelligent than its author. And indeed, it has found many important and fundamental applications. Firstly, the equation or variants thereof not only describe the electron, but in fact all elementary constituent particles like the leptons (electrons, muons, neutrinos etc.), and the quarks. Not so surprising as all of them have spin one-half and are fermions. Secondly, the Dirac equation and the field theoretic concepts that come with it are also extremely relevant

[6] Dirac himself hoped initially that the positively charged particle would correspond to the proton, so that the equation would somehow describe the complete hydrogen atom. It was Robert Oppenheimer, then at Princeton University, who pointed out that the oppositely charged particles had to have the same mass and therefore the equation implied a new species of particles, now denoted as *anti-matter*.

in condensed matter physics. This is somewhat surprising because *a priori* that is at low energy and you would not expect excitations to satisfy a relativistic equation. Yet, in a conductor the electrons fill the available energy states up to a level which is called the Fermi level. And there you have a situation which is like the Dirac vacuum, and indeed you can excite an electron which means that you effectively create a particle-hole pair.

Majorana fermions. There are even closer analogies, since quite recently so-called *topological phases of matter* have been predicted in which boundary excitations occur that are effectively behaving like massless Dirac modes, thus behaving like relativistic particles. An example is the so-called Majorana fermion, which is a special case where the particle is its own antiparticle. So it has only two components. The theory of the *Majorana fermion* goes back to thirties of the twentieth century, to a brilliant young Italian physicist who proposed the model, but then mysteriously disappeared. In fact his disappearance has never been fully resolved or explained. Whereas his person remains a mystery, his equation fortunately does not.

The mathematics of the Dirac operator. Finally, the notion of the *Dirac operator*, which is the first-order differential operator that defines the Dirac equation, plays an important role in pure mathematics. For example the index of the massless Dirac operator on smooth curved manifolds is directly linked to certain topological invariants of that manifold, through the so-called *Atiyah–Singer index theorem*. We will return to the Dirac equation in somewhat more detail in the next Volume.

Quantum Electrodynamics: QED

Quantum Electrodynamics is the first and very successful example of a quantum field theory. We outline some of its basic structure and properties, and mention states, operators and Feynman diagrams. This theory, starting from first principles, made some impressive, precise predictions that agreed with experiment up to 12 significant digits!

The first milestone in relativistic field theory was the formulation of *Quantum Electrodynamics (QED)*, a completely consistent quantum theory of electrons, positrons, photons and their interactions. The theory was completed just after the Second World War, quite independently, by the American physicists Richard Feynman and Julian Schwinger, as well as the Japanese Sin-Itiro Tomonaga. They jointly received the 1965 Nobel prize in Physics for this work. This success generated further developments in field theory which during 1970s culminated in the formulation of the successful *Standard Model* of all the known elementary particles and the fundamental forces between them.

Particles and force fields. In classical physics there is a clear (ontological) distinction between, on the one hand, constituent particles carrying mass and charge (like electrons and protons), that are often considered 'point-like', and on the other hand the force fields through which they interact like the electromagnetic field, and which spread out over all of space-time. In relativistic quantum field theory this distinction disappears. Particles correspond to ' wavefunctions' or states of quantum fields which can be spread out. But the arrow goes both ways, so classical force fields (like the electromagnetic field) when quantized have particle-like excitations (like the photon). And we say that the forces are carried or mediated by those particles. Particle-wave duality is lifted to a particle-field duality at a higher (or should I say, deeper) level.

So, the electron and its anti-particle the positron are described by a Dirac-type quantum field, as are the neutrinos and the quarks. A state of the electron quantum field may describe any number of electrons and/or positrons. So, there is *one* field for all electrons. In fact, every particle type has its own quantum field. But, also the force

fields of the strong and weak interactions have their own quantum fields, whose particle-like quanta we call *gluons* and *W and Z bosons*, respectively. In other words, there is a universal particle-field correspondence on the quantum level if we take relativity into account. Quantum field theory transcends the distinction between particles and forces in an essential way, yet the quantum fields describing constituent particles and force fields have distinctive features, because the constituents are fermions with spin one-half, and the force fields are bosons with spin one.

States and operators. A distinguishing feature of the quantum field theory framework is that it allows for the creation and annihilation of particles. Let me try to give a flavor of how that works. The first important ingredient is the existence of a vacuum or a ground state denoted as the zero state $|0\rangle$, that is the state without any particle in it. The second ingredient is that quantum fields can be expressed in terms of *particle creation and annihilation operators* that can act on the vacuum, or any other state, and create or annihilate a particle in that state. A generic state is in fact a multi-particle state that is labeled by the number of particles present in the state and what their energy, momentum and spin-polarizations are. For example a state,

$$|n_\gamma(\epsilon^\mu, k^\mu), n_e(s, p^\mu), n_p(s', p'^\mu)\rangle$$

would correspond to a state with n_γ photons in a state with four-momentum k^μ and polarization vector ϵ^μ, and so on. The electrons and positrons have spin one-half, and their spin-polarization is encoded in the variables $s\,(s')$.

Particle creation and annihilation.[7]

The physics we want to describe involves the creation and annihilation of particles and this is implemented by creation and annihilation operators we just mentioned. The photon field, for example, corresponding to the vector potential $A^\mu(\mathbf{x}, t)$, has a linear expansion in photon creation and annihilation operators, which are denoted $a^\dagger(\epsilon^\sigma, k^\nu)^\dagger$ and $a(\epsilon^\sigma, k^\nu)$. If the creation operator acts on a state, it creates a particle in the corresponding state, so for example:

$$a^\dagger(\epsilon^\sigma, q^\nu)\,|0\rangle = |n_\gamma(\epsilon^\sigma, q^\nu) = 1\rangle$$

Here the creation operator acts on the vacuum and creates a new state in which there is one photon present (n=1), with the specified polarization and energy-momentum. If you apply the annihilation operator to the vacuum, you would simply get zero:

$$a(\epsilon^\sigma, k^\nu)\,|0\rangle = 0\,,$$

because there is no particle to be annihilated. If there had been a particle with the corresponding properties in the state, that particle would be annihilated and we would end up with the vacuum state. But if we act on the vacuum state there is no particle to annihilate and the result is the number zero – the operator 'annihilates the vacuum' is the jargon.

In general one considers rather elementary processes, with a few incoming particles creating an incoming state, then these particles interact with each other (so typically particles will be annihilated and created), and what we want to know is what the possible final states are and what the probabilities are that they occur. To do these calculations, demands a lot of skill, since they tend to be extensive and it takes even the largest computers days to do the job. But the hardest part is also to set up the calculation and figure out in all detail which sub-processes will be there, and how important they are. It involves also an incredible amount of book keeping which of course has to be performed impeccably, and one therefore has to build in all kinds of checks and balances to see whether the extremely rigid laws are completely obeyed at any stage of the computation. Experiments like those at CERN are also at the forefront of all kinds of AI applications, both on the data analysis side

[7] I have had the pleasure of running into *creationists* and *nihilists*, but so far not into any *annihilists*.

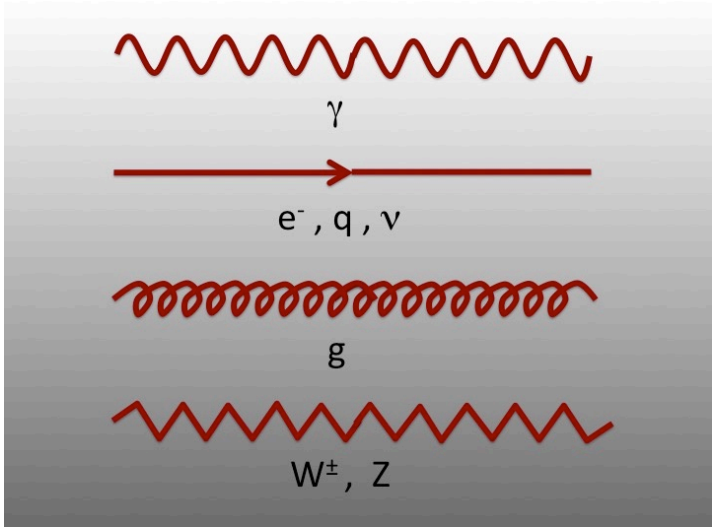


Figure I.4.28: *Propagators.* Particle propagation lines or *propagators* for various particle types. The arrow on the fermion line keeps track of the charge or better of the particle versus anti-particle degrees of freedom of the field.

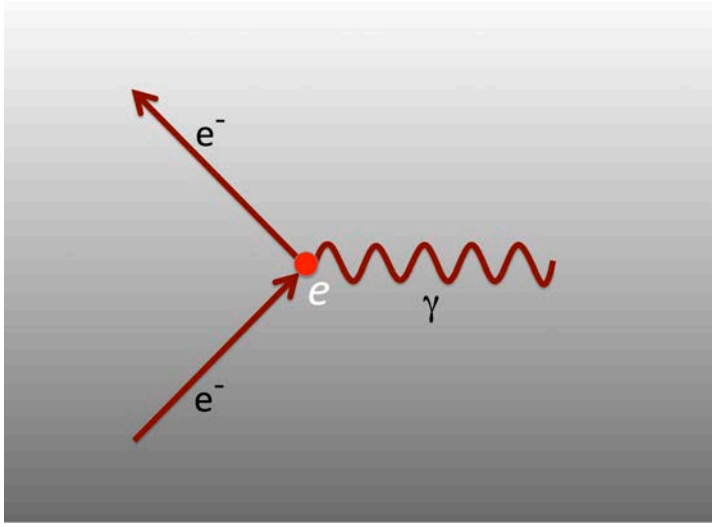


Figure I.4.29: *Interaction vertex.* The unique QED interaction is given by the interaction vertex of a photon with a charged particle, like an electron or quark. The strength of the coupling equals the coupling constant '*e*'.

as on the theoretical, calculational side, where we have to distinguish the numerical methods from the highly automated algebraic manipulation technology.

The language of Feynman diagrams. At this point the diagrammatic language created by Feynman becomes an indispensable tool. Let me give you an impression of how this methodology works. We have mentioned the (in- and out-coming) states, and these are represented as lines entering or leaving the diagram, where each particle type has its own type of 'propagation' line as illustrated in Figure I.4.28. The interactions are represented by diagrams where the particles that interact come together at a *vertex*. For example in Figure I.4.29 we see an electron emitting or absorbing a photon, where the electron moves on but with a different momentum.

The theory is relativistically invariant which means that you can make space-time 'rotations' or Lorentz transformations. This implies that you can also rotate the diagram clockwise over 90 degrees (as in Figure I.4.30 on the left), and you get the diagram for a photon coming in and an electron coming out and – help – what is that? It looks like an electron moving backwards in time! You may think so, but that is indeed what a positron is. A negative charge (electron) moving backward in time is the same as a positive charge moving forward in time, because that is the way the Dirac equation works. At the vertex – the red dot, the interaction takes place with a strength of the charge e , and in the interaction the energy, momentum and charge have to be conserved. So, what goes in, has in some form to come out again. If you had rotated the diagram counterclockwise instead (as in the same figure on the right), you would have obtained a diagram representing electron-positron annihilation into a photon, and indeed the total charge is zero at all times.

As far as QED is concerned these are roughly the fundamental rules but the diagrams may become arbitrarily complicated.

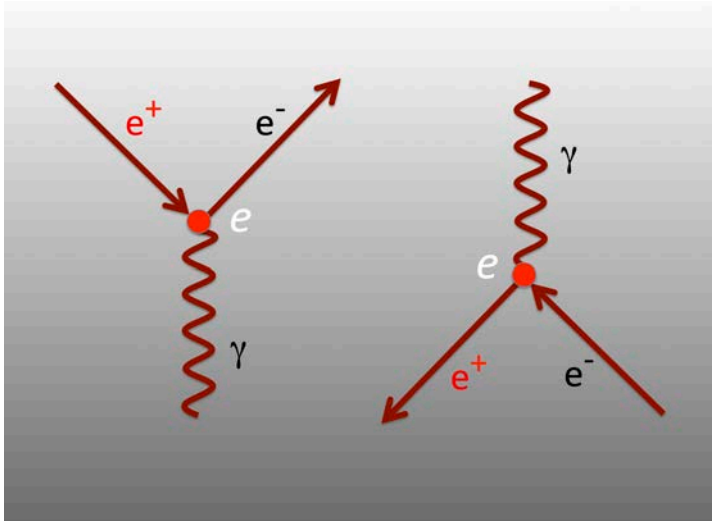


Figure I.4.30: *Interaction vertex.* The rotated diagram gives the coupling for the creation of an electron-positron pair on the left, and the annihilation of an electron-positron pair on the right. The convention is that in the diagram time goes upward. Note that the electron (negative charge) moving backward in time is the same as a positron (positive charge) moving forward in time.

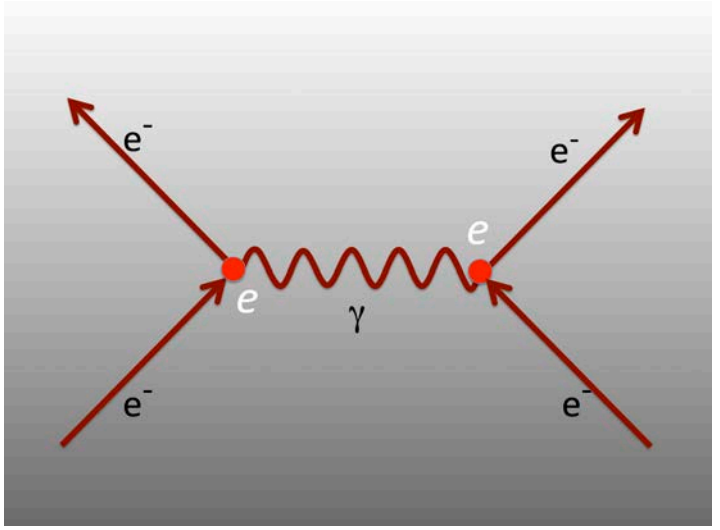


Figure I.4.31: *Photon exchange.* A lowest order photon exchange diagram contributing to the amplitude for electron-electron scattering.

That is the quantessence of the trade: you know what comes in and presumably also what comes out, and then, in principle, you have to construct all possible diagrams that – obeying the rules – can be drawn in between. You can of course order the diagrams by the number of vertices they have and if the coupling is small, then the contribution of higher-order diagrams becomes ever smaller. So, you stop after a few orders and get a sufficiently accurate result. You calculate the diagrams one by one and then add the results to obtain what is called the total quantum *probability amplitude* for the process.

As the word probability amplitude suggests, you have to square this expression to obtain the probability for the process to take place. In Figure I.4.31 we for example give one the (two) leading, lowest order diagrams that contribute to the probability amplitude for electron-electron scattering. What I am trying to convey is that the diagrams furnish a powerful and precise symbolic language which represents an intricate mathematical structure. The Feynman rules give you the unique translation of the diagrams into complicated but very explicit mathematical expressions that then have to be evaluated (mostly by computer) to get the real probabilities out.

Doing precision measurements means doing critical precision tests on theoretical models and that is the core of empirical science. Realistic precision calculations may involve hundreds or thousands of diagrams, and so even the generation of all the allowed diagrams is done by computer. In this field a lot of pioneering work in symbolic manipulation by computers has been done. Nobel laureate Martinus Veltman was the first with his program named '*Schoonschip*' which literally means 'clean ship', though in Dutch it actually means 'cleaning up the mess.' This program has gone through many upgrades and extensions and is still a program used by many practitioners. The most well-known outcome of such physics inspired artificial intelligent systems is the magnificent and versatile

symbolic manipulation and graphics platform *Mathematica* developed by Stephen Wolfram.

Subnuclear structure

The Standard Model

The Standard Model is a theoretical model for the basic constituent particles of all ordinary (meaning, not-dark) matter. Think of the leptons *such as the electrons and neutrinos, and the* quarks *that make up the protons and neutrons, but think also of the force-carrying particles that bind the constituents together. The model gives a unified description of three of the four fundamental forces: the electromagnetic, and the weak and strong nuclear forces. Gravity, however, is not included in the Standard Model. The model has made numerous precise predictions that so far have been vindicated by a variety of large-scale experiments in the world's biggest accelerators.*

The Standard Model was completed in the early 1970s. The experimental verification of many of its predictions took another forty years and still continues. A landmark was the discovery of the W and Z bosons at CERN (and somewhat later at Fermilab) in 1983. Another highlight was the discovery of the Higgs particle at CERN as recent as 2012. It was the last missing entry in the particle table of the model. The Higgs particle is a unique ingredient because it provided the explanation for the mass of other particles, in particular the masses of the weak force carrying W and Z particles. The presence of these masses is reflected in the fact that the corresponding interactions are short range as we discussed in the section on nuclear potentials and the Yukawa potential on page 166.

Figure I.4.32: This work of the Belgian surrealist painter René Magritte is entitled *Les Jeunes Amours* (1963). A more prosaic title, well fitting our sub-nuclear narrative would have been *A Color Triplet of Apple Quarks*. There is even a Dirac sea in the background. (Source: ©'Photothèque Magritte / Adagp Images, Paris)

Flavors, colors and families

To understand the structure of the Standard Model, let us look at Figure I.4.35, and explain what information is encoded in the colorful tables. In each of the figures, the top panel contains the force-mediating particles and the Higgs particle, these are all bosons, i.e. they have an integral spin. We shortly describe these in detail but it is more convenient to first turn to the content of the lower panels.

Particle families. The lower panels list the constituent particles: these are all fermions, and have spin one-half. There are three families of constituent particles denoted by three different colors as depicted in Figure I.4.35(b). Only the first family is stable, it consists of the *up* and *down*

quarks and the *electron* and its *neutrino*. These are the building blocks that make up all forms of stable (ordinary) matter in our universe. The other families consist of heavier but unstable copies of the light family, and sure enough they also play a crucial, albeit more hidden, role in our universe, such as processes inside stars or in the early universe. We first look more closely at the quarks and thereafter at the leptons.

Quarks: flavors and colors. The first thing to note about quarks is that they carry fractional electric charges. Those in the left column have a charge $+2e/3$, whereas the ones in the right column have $-e/3$, so that the proton which corresponds to two 'ups' and one 'down' has indeed a charge e , while the neutron made up from one 'up' and two 'downs' has zero charge. Besides their spin and charge, we distinguish two other intrinsic properties that were briefly mentioned before: *flavor* and *color*.

Flavors. The so-called *flavor* index corresponds to one of the six letters $(\mathrm{u}, \mathrm{d}, \mathrm{s}, \mathrm{c}, \mathrm{t}, \mathrm{b})$ ′ which in turn refers to their names *up*, *down, strange, charm, top* and *bottom*.

Besides the lightest nuclear particles or hadrons like the proton, neutron that make up stable matter, there are many nuclear particles that also involve quarks of flavors other than the up and down, but as those quarks are heavier, the particles in which they appear tend to be unstable. By the way, I always found the use of the word 'flavor' in this context a bit strange. What is 'up' or 'down' supposed to taste like, you wonder. The peculiar collection of flavor names for quarks has repeatedly given rise to exotic if not funny, even sexist expressions in titles of articles (involving topless or bottomless particle models etc.), which after submission were of course instantly refused by the editors of the established journals.

Flavor symmetry: the 'eightfold way.' From a historical point of view it is interesting to restrict ourselves to the three-flavour case. It is the case described by Gell-Mann

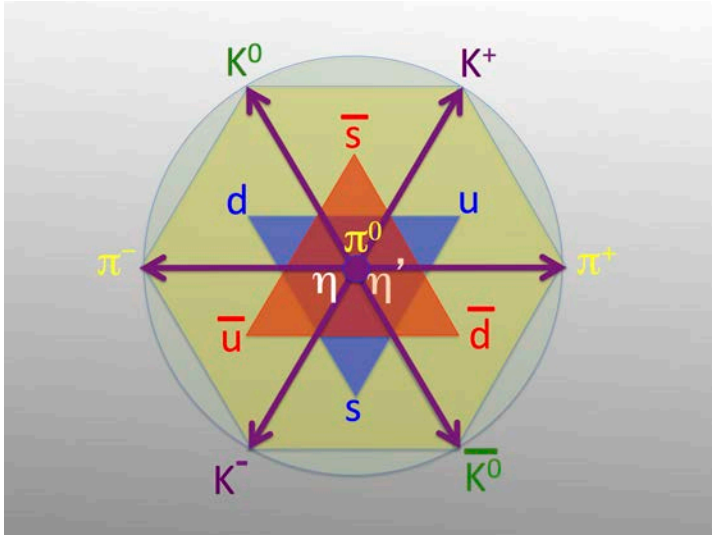


Figure I.4.33: *The* $\mathrm{SU}(3)$ *way.* Gell-Mann's eightfold way is based on the symmetry group that classifies the *flavor* properties of the particles making up the 'particle zoo.' The geometric patterns by which the particles are labeled actually have a three- or sixfold symmetry rather than an eightfold one. The observed particles are the ones on the outer hexagon and the three particles at the origin, together they form the *meson nonet*. The symmetry has two fundamental three-dimensional representations corresponding to the two triangles in the center, which suggested the existence of three quarks $(\mathrm{u}, \mathrm{d}, \mathrm{s})$ and their anti-particles $(\bar{\mathrm{u}}, \bar{\mathrm{d}}, \bar{\mathrm{s}})$.

in his 'eightfold way', based on a $\mathrm{SU}(3)$ flavor symmetry group. To give you a flavor we have depicted some of the geometric representations in which the particles are classified according to this $\mathrm{SU}(3)$ scheme in Figure I.4.33. This representation is called the *meson nonet*, referring to the nine possible quark anti-quark combinations of the up, down and strange flavors, which gives $3 \times 3 = 9$ combinations. This representation is one of the examples that makes up the aforementioned 'particle zoo' of nuclear particle states. The fundamental particles form a triplet representation of quarks and an anti-triplet of anti-quarks, and these correspond to the blue and red triangles in the center.

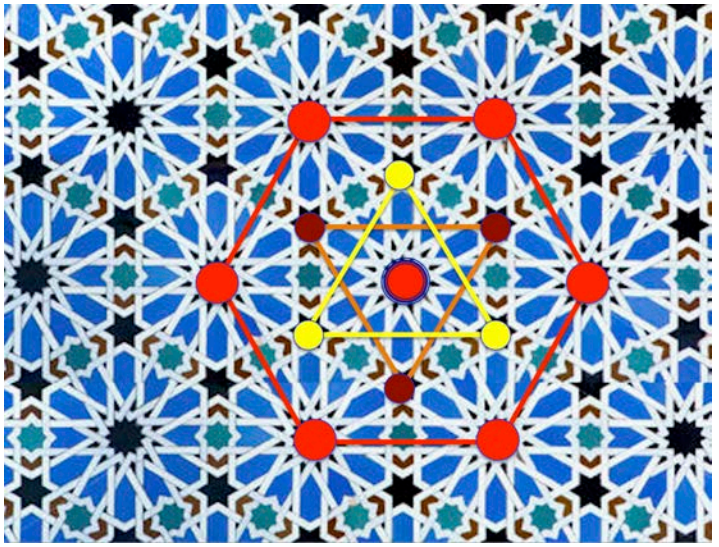


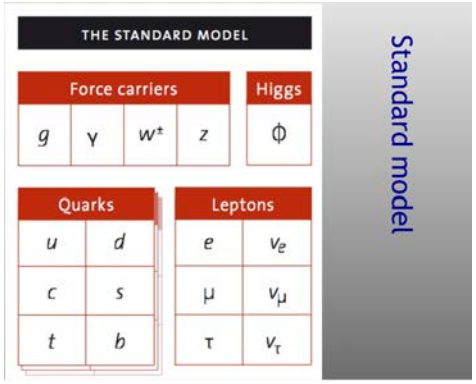
Figure I.4.34: *Quarks and* $\mathrm{SU}(3)$. A beautiful early Islamic tiling from the *Real Alcázar* (Royal Palace) in Sevilla in Spain, exhibiting the sixfold symmetry characteristic for the group $\mathrm{SU}(3)$. The black stars represent the weight-lattice of states corresponding to the group $\mathrm{SU}(3)$ The *representations* correspond to certain triangular or hexagonal subsets of states centered at the origin.

Quarks have never been observed as individual, freely moving particles; they are confined to composites that consisted of quark anti-quark pairs (the mesons) or (anti) quark triples (the baryons). The nonet consists of nine mesons made up of one of the three quarks paired with one of the three anti-quarks in the figure. These scalar (spin zero) particles, are, as you see, siblings of the pions we have mentioned before. The three particles in the center correspond to different linear combinations of the $\bar{\mathrm{u}}\mathrm{u}$, $\bar{\mathrm{d}}\mathrm{d}$ and $\bar{\mathrm{s}}\mathrm{s}$ pairs. Why the 'eightfold way' you are inclined to ask, while the picture clearly exhibits a sixfold symmetry? It turns out that eight of the nine mesons basically form an *octet*, that is a larger irreducible representation of the group, meaning that under the $\mathrm{SU}(3)$ transformations those particle states would be transformed into each other. The ninth (η') particle is all by itself and invariant under the symmetry group. In the era when this scheme was proposed all the observed particles could be catalogued in certain $\mathrm{SU}(3)$ representations (like the octet mentioned before), and this of course shifted the quest for fundamental building blocks to the underlying level of the quarks.

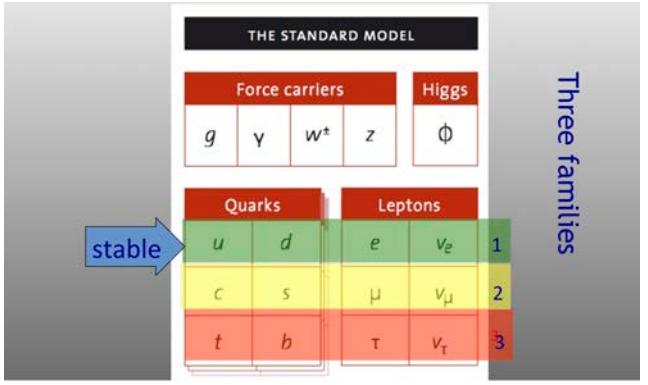
If the symmetry was exact, then that would imply that the baryons or mesons that belong to a single representation of the symmetry group should have the same mass; the particles should be degenerate. This turns out not to be the case here, and therefore we say that flavor is only an approximate symmetry. Nevertheless having the symmetry patterns and the observed particles and their masses, Gell-Mann could see that certain particles were missing from the observations, and that way he could make quite precise predictions of their properties and therefore also say where to look for them. An example is the Ω^- particle belonging to the decuplet representation of baryons and discovered in 1964. This is indeed reminiscent of the story of Mendeleev and his periodic table.

It is actually somewhat ironic that the $\mathrm{SU}(3)$ or the much larger $\mathrm{SU}(6)$ flavor symmetry does not really feature in the Standard Model as we see from the panels in the figure. The flavors are there but they come in pairs, which refers to the weak interactions as we will explain shortly. The 'eightfold way' is completely accidental from the Standard Model perspective. Nevertheless the family structure is very much present and even required for the consistency of the model. But that family structure as such is not *explained* by the model. It is one of the challenges to look for yet more involved schemes.

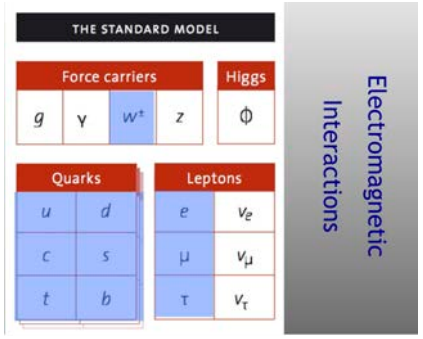
Color. The second property of quarks refers to what is called their *color*. Each flavor comes in three different '*colors*', usually denoted as *red, green* and *blue*. In the figure that is visualized by the stack of three quark panels on top of each other. This 'color' quantum number is some kind of charge to which the strong nuclear force couples, and needless to say, has nothing to do with ordinary color. This nomenclature is at least consistent, which cannot be



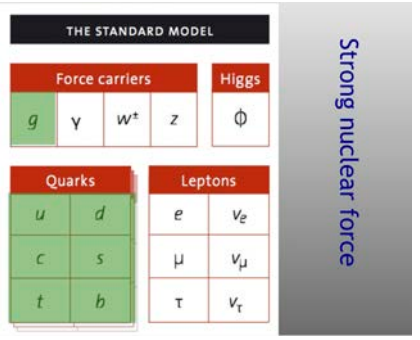
(a) The constituent particles, quarks and leptons are fermions. The particles mediating the forces are bosons, and the Higgs boson generates mass for the particles.



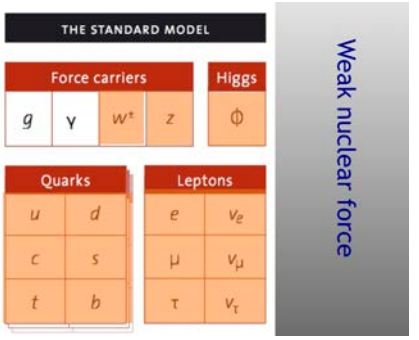
(b) The constituent particles come in three families. All ordinary matter is made of the first family of lightest and therefore stable particles.



(c) The electromagnetic force mediated by the photon affects all particles that carry electric charge.



(d) The strong nuclear force only mediates between the three different colors of quarks and does not distinguish flavor. It binds quarks into color neutral nuclear particles.



(e) The weak nuclear force affects all constituents, but it does not mix quarks with leptons within a family.

Figure I.4.35: *The standard model.* Constituent particles and how the basic forces act between them.

said of the flavors.

The excellent 'artist impression' of flavor and color properties of quarks is given by the Magritte painting of Figure I.4.32. The painting dates from 1963, one year before Gell-Mann and Zweig proposed the existence of quarks, but well before the color property of quarks was postulated, implying that the quite striking correspondence is entirely coincidental.

Leptons: electrons and neutrinos. In the bottom panels on the right in Figure I.4.35 , we have listed the three families of leptons: the *electron*, the *muon*, and the *tau* family, including their respective neutrinos. The neutrinos have pretty ghostly properties in that they have no charge and were long believed to be massless. It has quite recently been established, however, that they have tiny masses. They only interact weakly (and gravitationally), which means that we don't see or feel them, in spite of the fact that we are permanently bombarded by billions of these neutrinos per second. They basically fly unhindered through most things, like the Earth for example. The evidence for their existence was for a long time just based on their absence, since the amount of missing energy and momentum in weak-decay processes pointed to the existence of a massless, neutral particle – a neutrino therefore. A tiny brother of the neutron. To catch a few of them we have to build detectors consisting of an incredible number of steel plates with very special (so-called flat wire chamber) detectors in between, and that is how after a long time their existence was established in a direct fashion. The electron neutrino was the first to be discovered in 1956 by Frederick Reines and his collaborators. He shared the 1995 Nobel prize for Physics for this discovery with Martin Perl who discovered the tau-neutrino in 1974, quite some time after Leon Lederman, Melvin Schwartz and Jack Steinberger received the prize for the muon-neutrino in 1988.

The matching of the lepton and quark panels in the figures is essential for the consistency of the model. But it is not known whether the family structure can be explained by some underlying mechanism, where the different family levels are excited levels of some underlying structure.

Force mediators. In the top panels we see the force mediating particles and the Higgs particle. The force carriers have spin one, which means that they are vectors like the electromagnetic gauge potentials. The Higgs has spin zero; it is a *scalar* particle without spin degree of freedom. To see what interactions these force particles mediate, it is best to look at the three figures at the bottom. On the left in Figure I.4.35(c), we have the familiar electromagnetic interaction mediated by the *photon* denoted by γ , which is described by the QED part of the Standard Model. Electromagnetism only affects the blue-colored particles, which are the particles that carry electric charge. Note, that all constituent particles carry charge except the neutrinos. And that is exactly the reason we can't observe them very easily. The force particles (including the photon itself) are electrically neutral except for the $W^{\pm}$ particles which carry a unit of charge. In the middle figure, we display the strong nuclear force which only works between quarks mediated by the *gluons* denoted by g . This brings us to the theory of the strong interactions to which we now turn. We will discuss the weak interactions of Figure I.4.35(e) in more detail thereafter.

The strong interactions

Quantum chromodynamics (QCD). The quantum theory for the strong nuclear force is called *Quantum Chromodynamics (QCD)*. The strong force is mediated by 8 *gluons*, which are described by 8 color gauge potentials, that manifest themselves in the presence of 8 'color-electric' and 'color-magnetic fields.'

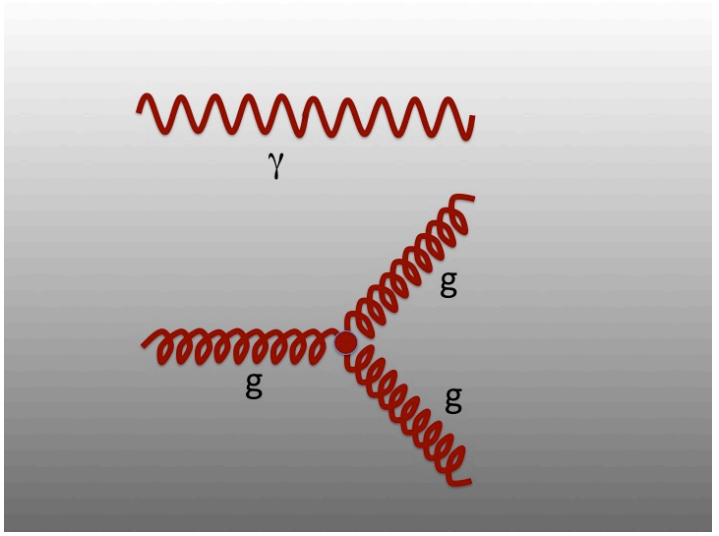


Figure I.4.36: *Self-interaction of gluons.* The photon (γ) has no self-interaction. The gluons (g) have self-interactions, which makes the theory very much nonlinear and much harder to deal with. The three gluon vertex is represented differently on the left of the next figure.

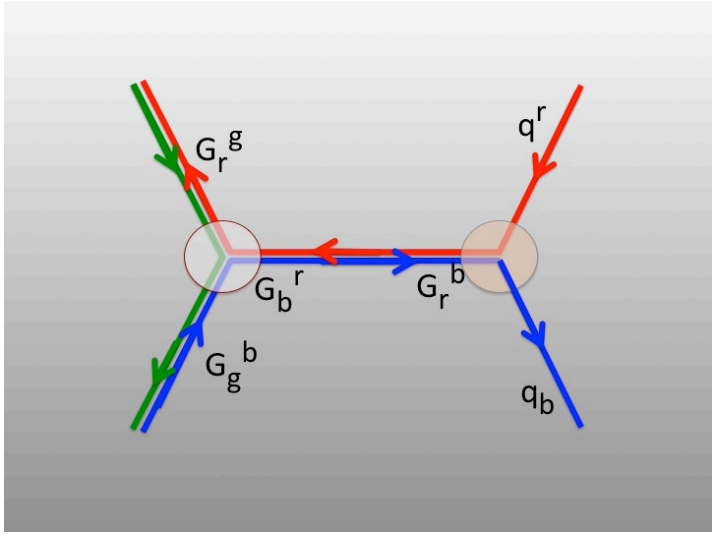


Figure I.4.37: *Color-flow diagram in QCD.* A nice way to visualize the interactions in QCD. Quarks carry a single color line, while gluons carry two (different) lines. In the vertices the color charge is conserved, so, the colors and arrows have to match. Upper index goes into the vertex, lower index goes out.

Self-interactions. A crucial difference with QED is that the gluons themselves also carry color charge like the quarks. This means that they interact with themselves and that understandably leads to complex nonlinear behavior. The reason why electromagnetism is so much simpler is precisely that the photon does not carry electric charge and therefore does not interact with itself. This means that pure electromagnetism without charges and currents is a linear theory and indeed the source-free Maxwell equations are linear as we saw in Chapter I.1. These have simple sinusoidal wave-like solutions, which on a quantum level correspond to freely propagating photons. The essential difference between the non-self-interacting photon and a self-interacting gluon is indicated in Figure I.4.36. In addition, the effective strength of this color coupling is large, so it is hard to make successive approximations to higher order in the coupling. The language of Feynman diagrams loses much of its power because it is an approximation scheme that involves successive powers of the coupling constant. If that coupling is small the series is expected to converge and it suffices to only keep a limited number of lower order contributions to obtain a meaningful result. If that coupling becomes large the successive contributions keep increasing and one loses the convergence and hence the ability to make meaningful calculations and reliable predictions.

An alternative way to think about gluons, quarks and the way they interact with one another, is given in Figure I.4.37, where the (anti-)quarks are denoted by a single directed color line, and the gluons as an oppositely directed pair of lines. The picture is illuminating in that it shows very clearly what it means to say that color (charge) is locally conserved. The figure is not meant to imply that the gluons are actually made up of (anti-)quarks. Though they can manifest themselves in the same color anti-color 'channel', the gluons represent independent physical degrees of freedom. The fact that strong self-interactions lead to

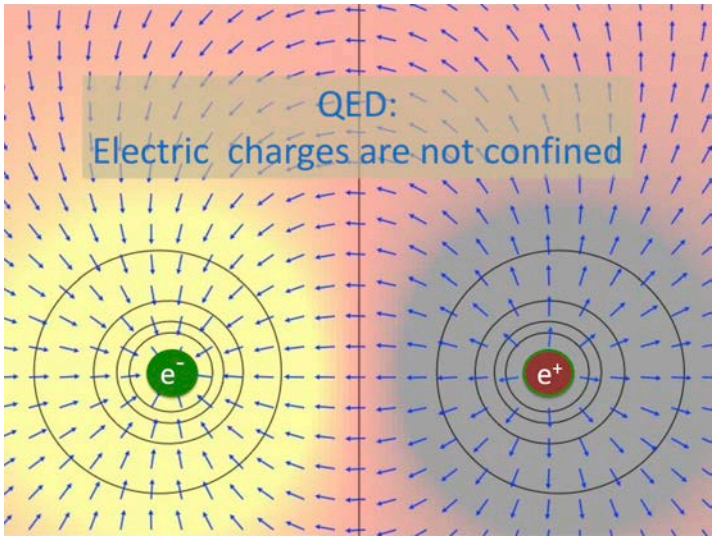


Figure I.4.38: *Free electric charges.* The electric field lines connecting the arrows go from the positron to the electron but they spread out widely over space. The electric charges are therefore not confined and we can observe them as free particles.

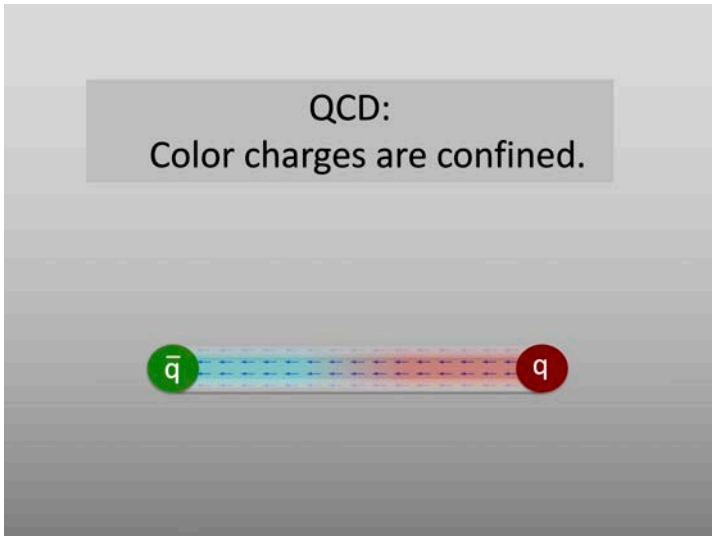


Figure I.4.39: *Confined color charges.* In QCD the color electric fields do not spread but are forced into a narrow tube which leads to the *confinement* quarks. It is a consequence of the highly non-trivial nature of the ground state of QCD which behaves like a color magnetic superconductor.

unexpected behavior may not sound unfamiliar to us humans. Anyway, it is this feature in QCD that made it so hard to see from the basic structure of the theory what the resulting physics and phenomenology of quarks and gluons would be.

Confinement. The binding mechanism between two quarks is very different from the attraction between to opposite electric charges. This is illustrated in Figures I.4.38 and I.4.39. In the first figure we see that the electric field between two opposite charges spreads over all of space, reflecting the $1/r^2$ force law. It is as if the field lines repel each other. In this case we can give one of the charges enough energy that the pair breaks up into two free charges. The second figure shows what the color-electric fields between a quark and an anti-quark look like. The field lines are squeezed into a narrow tube that connects the pair. It is as if the field lines attract each another. The energy per unit length of the electric flux tube is constant because the tube is everywhere the same. This in turn implies that the interaction energy of the pair grows linearly with their separation. It would increase indefinitely if not the energy at a certain point exceeds the energy needed to create a new quark anti-quark pair somewhere in between. Then what we basically have done is to create two pairs out of one pair! We cannot create a separate quark. because as a source it always has to stay connected to a tube.

More in general it turns out that the color force works in such a way that only color neutral composites of quarks, denoted as '*color singlets*,' can exist as free particles. And therefore these are the nuclear particles that we observe in nature. The way this usually is expressed is to say that color is *confined*. The property of color is hidden. Quarks and gluons are for ever *imprisoned*. The simplest singlets are either made-up of three quarks with different colors (these are the *baryons* like the proton and neutron mentioned before), or of color anti-color quark pairs (the

mesons like the pions). And the same is true for the gluons themselves, because they carry color charge; they only appear in color neutral composites which are called *glue balls.* It is this constraint of color neutrality that explains quarks and gluons are confined and why we cannot observe them as free individual particles like electrons.

The confinement phenomena presents us with a unique, paradoxical situation we did not encounter before. QCD is a theory formulated in terms of fundamental physical degrees of freedom (quarks and gluons) that are not discernible, so that from a philosophical point of view you are tempted to question their very existence. 'To be or not to be, that is the question!'

Asymptotic freedom: how strong becomes weak. What does it mean to say that interactions are strong? In this case it is a relative statement in that it is a force between protons and neutrons, or on a more basic level between quarks, that is strong enough to overcome the Coulomb repulsion so as to make nuclear binding possible. This implies that the coupling strength of the interaction is considerably larger than that of electromagnetism. What we mean to say is that the effective dimensionless number characterizing the strength of the interactions must be much larger, and this amounts to saying that the analogue of the electromagnetic fine-structure constant $\alpha = e^2/(4\pi\hbar c) \simeq 1/137)$, which for the strong interactions is called α_s , is of order unity. This tells us that at the relevant nuclear scale of 1 $\mathrm{fermi} = 10^{-15}$ m the effective coupling is large.

The confinement picture I.4.39 shows that the color fields emanating from the quark are forced into a narrow tube that terminates at some antiquark. The tube has a cross-section which is typically of the confinement scale, say, one fermi squared. So here is how we should think about this. For distances much larger than one fermi, the quarks are confined, which means that the complicated nonlinear self-interactions of the gluons have collectively created an effective environment that causes the confinement. However, for distances much smaller than one fermi the quarks are effective moving 'freely,' in the sense that the effects of the self-interactions are negligible. In fact on such scales one could treat the strong interactions more like a type of electromagnetic interactions. The color field lines go radially out of the quark and bend over in the confining tube at a distance of about one fermi. On that small scale one could use the perturbative approach in terms of Feynman diagrams to calculate the dynamics.

It is interesting to look at the result of such intricate calculations of the effective coupling strength as a function of momentum transfer (or inverse distance) both for α (QED) and α_s (QCD). For QED one obtains,

$$\alpha(q^2) = \frac{\alpha}{[1 - (\alpha/3\pi)\ln(q^2/m^2)]} \quad \text{for} \quad q^2 \gg m^2 \,, \tag{I.4.30}$$

where $\alpha = 1/137$ and m denotes the relevant mass scale one is interested in. For QCD one obtains,

$$\alpha_s(q^2) = \frac{12\pi}{(33 - 2f)\ln(q^2/\Lambda^2)} \quad \text{for} \quad q^2 \geqslant \Lambda^2 \tag{I.4.31}$$

where $f = 6$ equals the number of flavors and Λ sets a mass scale at which one is interested. We have plotted these curves in Figure I.4.40. There are two striking differences between the two curves: (i) the relative difference in strength on the scales we are interested in is indeed big, about a factor of one hundred, and (ii) the strong interaction is decaying substantially for increasing momentum and thus for smaller distances. *The strong interaction gets weak at small distances!* This property is called *asymptotic freedom.* It is of crucial importance because it allows for precise calculations of high-energy scattering processes where you probe very small distances and compare those to the experiments.

So what happens if two quarks collide head-on in a collider? They may strongly scatter and the outgoing quarks or gluons may get a high transverse momentum. These

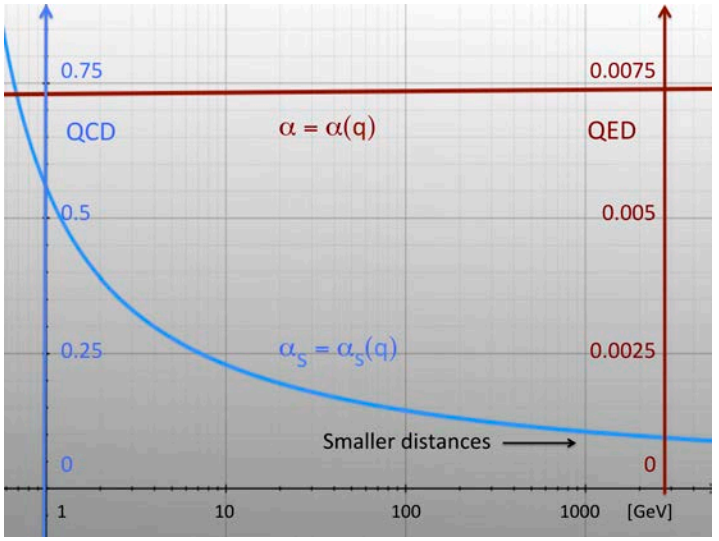


Figure I.4.40: *Asymptotic freedom.* Plots of the effective coupling strength as a function of momentum (or probing distance) for QED and QCD. Note the scales differ by a factor one hundred. The fact that the blue QCD strength becomes weaker at short distances is called *asymptotic freedom*. It means that in the high-energy regime the theory can be studied by the diagrammatic (perturbation theoretic) method.

individually colored particles have to pick up companions to make color singlets at a scale of one fermi and will in the end cause a so-called *jet* of outgoing singlet particles. This highly collimated shower of particles has a total momentum equal to that of the originally scattered quark, so that individual quark momentum is an observable in the above sense.

We can also turn the story around and start at very small distances where the theory is very well behaved and our intuitions make sense because the system is weakly coupled. If we move up in scale towards the infrared the coupling becomes stronger, and when the coupling becomes of order unity the system becomes strongly coupled and our predictive ability breaks down. Now what this quite often means is that something drastic like a phase transition is going to happen. The ground state of the system becomes unstable and will change. For example a non-trivial condensate may form, and in fact in a sense the nature of the condensate in QCD is quite well understood and investigated (by computer simulations). The idea is that there is a condensate of magnetic degrees of freedom, monopoles and fluxes, so that the ground state of QCD is very much like a *magnetic superconductor*, a medium which would indeed confine color electric charged particles, like quarks and gluons.

To get some understanding of this mechanism we should look at ordinary (electromagnetic) superconductivity (type II) which will be discussed in Chapter III.3 in Volume III. The ground state of an type II superconductor corresponds to a condensate of electron (so-called Cooper) pairs. These cause the so-called *Meissner effect*, which means that magnetic fields are expulsed from the medium. If you turn on a strong magnetic field over a slab of superconducting material, then thin filaments of one unit of flux will penetrate the superconductor. Now imagine that I have magnetic monopoles to play around with and suppose that I drag that monopole into the superconductor, what would happen? Indeed, the magnetic flux of exactly one unit emanating from the monopole would be forced into one such narrow filaments and look for a way out at the boundary of the superconductor where the field would spread out again. But that is nothing but saying that monopoles would be *confined* in such a superconducting medium! And the dual of this mechanism is operative in QCD, a color-magnetic condensate confines the color-electrically charged quarks and gluons.

A final comment on this beautiful theory. Can we not in some way reformulate the theory in what we call a strong coupling regime where one over the coupling constant is the new coupling, which then can be taken to be small. This question was answered by Kenneth Wilson from Cornell University, and it amounted to a formulation of gauge theories on a discrete space-time lattice. in terms of link-variables like the ones we considered on page 35 in Chap-

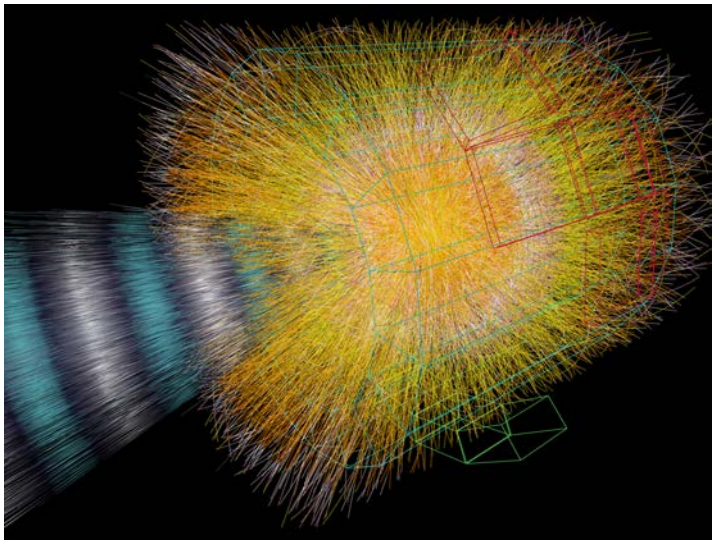


Figure I.4.41: *Lead-ion collisions.* A simulation of a lead-ion collision event for the ALICE detector at CERN, producing an enormous number of particles. To find what you are looking for is far worse than searching for a needle in a haystack! In these experiments one tries to recreate the conditions that were present throughout the universe shortly after the Big Bang.

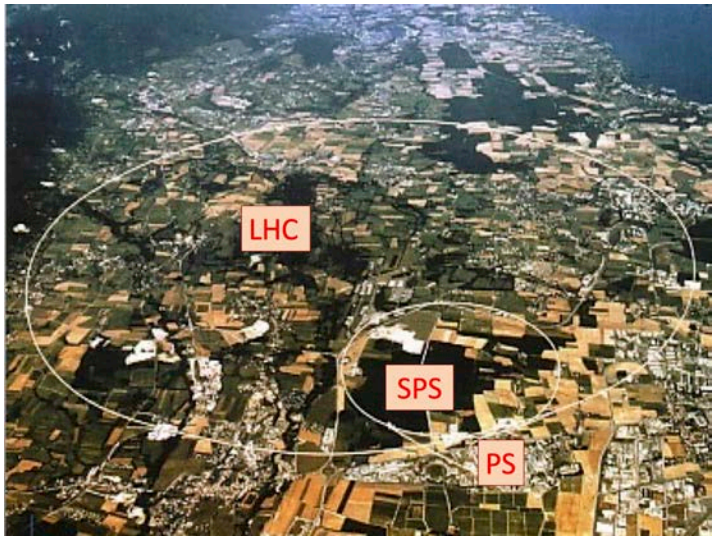


Figure I.4.42: *The Large Hadron Collider at CERN.* The largest accelerator at this moment is the Large Hadron Collider (LHC) at CERN in Geneva. The protons are accelerated in two oppositely directed circular beams. The circumference of the large ring is 27 km. Pre-acceleration happens in the older Proton Synchroton (PS) and the Super Proton Synchroton (SPS) accelerators.

ter I.2, where we discussed the line integral of a gauge potential. In this formulation one can make a very controlled and systematic strong coupling approximation to QCD, and the most immediate success is that confinement is there right from the start. This means that in this approach the lowest order calculation of the interaction energy between two external quarks yields a linear potential between them, and that is what confinement means. In Figure I.4.39 we see that the field energy per unit length is constant. Then the question became to prove that there was no discontinuity (a phase transition) between the weakly coupled and strongly coupled regimes. This turned out to be the case and with that the lattice approach to QCD has become an indispensable tool in the study of the strong interactions. Wilson was awarded the physics Nobel prize in 1982 for his profound work on phase transitions, which is embodied in his fundamental work on the *renormalization group*, a very general approach to studying the scaling properties of physical systems that we will return to in Chapter III.4. This work established a deep connection to the work on phase transitions in statistical and condensed matter physics by Michael E. Fisher and Leo Kadanov, and the renormalization program in quantum field theory going back to the early days of QED.

The quark-gluon plasma. If we shoot two protons with very high-energy onto each other, they surely break up, and what comes out are avalanches (called *jets*) of color-singlet particles – nuclear, but also leptons. Indeed in modern experiments the energies are so gigantic that thousands of particles are created in a single collision, as indicated in Figure I.4.41 showing (simulation of) a high energy event in the ALICE detector of CERN. In this experiment the physicists are trying to create a new high den-

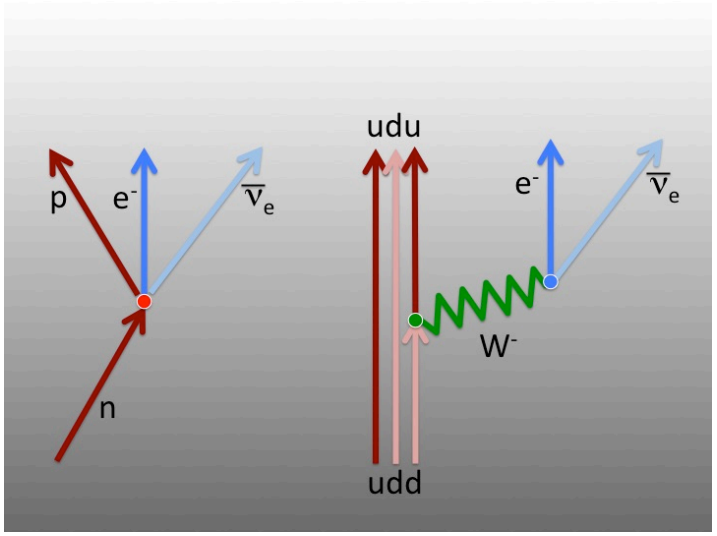


Figure I.4.43: *Beta-decay.* The diagrams for the beta-decay process. On the left a diagram at the level of nuclear physics, and on the right the resolution of the same diagram at the finer scale of the Standard Model. Note that the single vertex on the left corresponds to a pair on the right where the process is involving a W^- particle mediating the weak nuclear force.

.

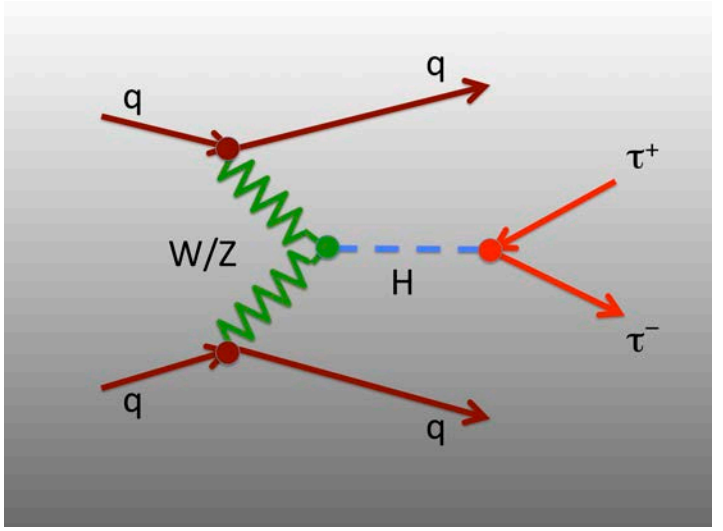


Figure I.4.44: *Higgs production.* A standard model diagram representing a particular process by which the Higgs particle is produced from the scattering of two quarks. The experimental signature of this process is provided by the two tau leptons in which Higgs instantly decays.

sity state of matter denoted as the '*quark-gluon plasma*', a state that may have existed in the very early universe directly after the Big Bang. They do that by banging leadions with very high energy into each other so that thousands of new particles are created, and for a fraction of a second these form a strongly interacting hot plasma made up of quarks and gluons with striking properties that should resemble the state of matter at the very early stages of the universe.

The electro-weak interactions

The W and Z particles. Let us return to the tables representing the Standard Model and to the elecrtro-weak interactions in particular. In Figure I.4.35(e) we focus on the weak nuclear force, mediated by the charged $W^{\pm}$ and neutral Z particles. It affects all constituent particles in an interesting way, the W bosons induce horizontal transitions in the table, because they are electrically charged. Their interaction vertices, allow for fundamental processes like:

$$u + W^- \rightarrow d\,, \tag{I.4.32}$$
$$e + W^+ \rightarrow \nu_e\,, \tag{I.4.33}$$
$$\mu + Z \rightarrow \mu\,. \tag{I.4.34}$$

The horizontal moves stay within the (color)panels, so red quarks to red quarks, electron to its neutrino and so on. A transition from a lepton to a quark is not possible because the W bosons have unit charge and that doesn't match the fractional difference in charge between a quark and a lepton. This in turn implies that the net number of quarks and the net number of leptons are separately conserved in these interactions. Take for example the process of 'β-decay' of the neutron where:

$$n \rightarrow p + e + \bar{\nu}_e\,, \tag{I.4.35}$$

as depicted on the left in Figure I.4.43. Recalling the compositions $\mathbf{n} = (\mathrm{udd})$ and $\mathbf{p} = (\mathrm{uud})$, the decay process above is in the Standard Model perspective composed of two non-trivial vertices on the constituent level:

$$\mathrm{d} \quad \rightarrow \quad \mathrm{u} + W^{-}\,, \tag{I.4.36}$$

$$W^{-} \quad \rightarrow \quad \mathrm{e} + \bar{\nu}_{\mathrm{e}}\,, \tag{I.4.37}$$

and this process is depicted on the right-hand side of the figure. The 'weakness' of the transitions comes about because the probability of creating an intermediate W and Z particle is very low due to their large mass. That results in an energy barrier which suppresses the transition process.

The Higgs particle. Finally, in Figure I.4.44, we show the complicated diagram of a process that contributes to the production of a Higgs particle (H) in the collision of two protons (or better, two quarks), where these exchange a weak W/Z boson, which can radiate from a Higgs. This Higgs is extremely short-lived and is not directly observed. The signature of the Higgs production in the out-coming state is the presence of two τ leptons. The Higgs was found in 2012 by two large international experimental collaborations: ATLAS and CMS in the Large Hadron Collider (LHC) at CERN.

The Higgs particle is an essential ingredient of the standard model as it is involved in a mechanism by which the masses of the W and Z particles are generated. This is discussed in more detail in Chapter II.6 on symmetries and their breaking.

This concludes our lightning review of what the cherished Standard Model of particle physics is about. In the next section we further explore the unification process in the successive formulations of fundamental physics at the subsequent stages of understanding.

A brief history of unification.

> There are two possible outcomes: if the result confirms the hypothesis, then you've made a measurement. If the result is contrary to the hypothesis, then you've made a discovery.
>
> *Enrico Fermi*

We have so far talked mainly about the fundamental building blocks and that translates into an inventory of what has been observed in experiments up to now. We have also reflected on the models for the interactions between these building blocks that account for the spectrum and the hierarchy of physical states. It is then interesting to step back and look at the history of theories, which is indeed also a history of concepts in theoretical physics. In Figure I.4.45 we have depicted this historical account focussing on the unification concept.

On the bottom line we list the basic classes of physical phenomena concerned, and going upward we also observe how they are linked to the fundamental forces, but we see also a progressing unification in the description of the fundamental physics. The two lines at the top represent theoretical developments which are still considered to be speculative and for which we eagerly await new experimental clues. This figure nicely illustrates the fundamental paradox of how ultimate *reductionism* may well lead to a form of ultimate *holism*!

Returning to the unification aspect, the first example is Newton's theory of gravitation (1687) that unified heavenly and terrestrial mechanics. Another beautiful example is provided by Maxwell's theory of electromagnetism (1865), which clearly unites electric and magnetic phenomena in one framework, but also includes electromagnetic fields and radiation like light, and therefore the subject of optics.

After the Second World War we learned to appreciate and include the quantum principles as in Quantum Electrodynamics (QED), the theory of photons, electrons and their anti-particles the positrons. This highly successful theory motivated theorists to find theories for the weak and strong nuclear forces based on similar principles.

The starting point was an approximate phenomenological model for the weak force proposed by Fermi in 1932. In the late 1960's this was replaced by a consistent unified quantum theory for both the electromagnetic and weak forces. Many names are actually connected to this development, firstly Sheldon Glashow, Abdus Salam and Steven Weinberg, who formulated the theory including the particle content including the weak force mediating particles. For this work they shared the Physics Nobel Prize in 1979.

This model was then augmented with the all-important ingredient of the *Higgs field* by Peter Higgs, Robert Brout and Francois Englert. Brout died in 2011, and therefore Higgs and Englert shared the Physics Nobel prize in 2013, shortly after the particle was discovered at CERN. Finally we should mention the seminal contributions of Gerard 't Hooft and Martinus Veltman, who constructed the consistent mathematical framework which enabled them to prove that the electro-weak theory was *renormalizable*, and which made comparisons of detailed predictions of the electro-weak theory with precision experiments possible. They received the 1999 Nobel prize for Physics for this work.

The developments for the strong interactions took place partly at the same time. Chen Ning Yang and Robert Mills proposed in 1954 the fundamental generalization of the Maxwell theory by extending the notion of the electromagnetic gauge invariance from the simple $\mathrm{U}(1)$ group to the non-abelian group $\mathrm{SU}(2)$. This led to a totally new, very beautiful non-linear system of equations, not surprisingly called the *Yang-Mills equations*. But it took quite some time before it was recognized that these equations formed the basis for the theories of both the strong and weak nuclear forces. In the section on Gauge symmetries of Chapter II.6, we discuss these symmetries and equations in more detail.

One of the leading scientists in the particle physics developments was the American Murray Gell-Mann who proposed the existence of quarks at the same time as but independently from George Zweig in 1964. Gell-Mann received the Physics Nobel prize for this and other contributions in 1969. After that he also formulated Quantum Chromodynamics (QCD) with his collaborators Heinrich Leutwyler from Switzerland and Harald Fritzsch from Germany in 1973. This theory is based on the Yang-Mills equations for the color gauge group $\mathrm{SU}(3)$. The binding mechanism and confinement of quarks was largely proposed by Yoichiro Nambu who received the Nobel prize in 2008. The property of QCD called *asymptotic freedom* made it possible to make sensible predictions for the strong interactions at high energies, This was discovered in 1973 by the American physicists David Gross, David Politzer and Frank Wilczek who received the Nobel prize for their work in 2004. We will say more about this shortly.

Forces of nature, Unite! Let me once more emphasize that the unification in the description of such a wide variety of physical phenomena in the Standard Model was possible because the different components are based on the same conceptual principles. These principles are those of *quantum theory*, those of *special relativity*, and the principle of *local gauge invariance*. The latter principle manifested itself in Maxwell's theory as we discussed in Chapters I.1 and I.2, in Einstein's general theory of relativity, and also in the Yang-Mills equations. Gauge invariance is a key ingredient because it is strongly tied-in with the notion of a force field and completely fixes what the interactions between the forces and particles look like.

In Figure I.4.45 you see that we have added two more rows on top. They express some powerful ideas leading to further unification, ideas that go beyond the Stan-

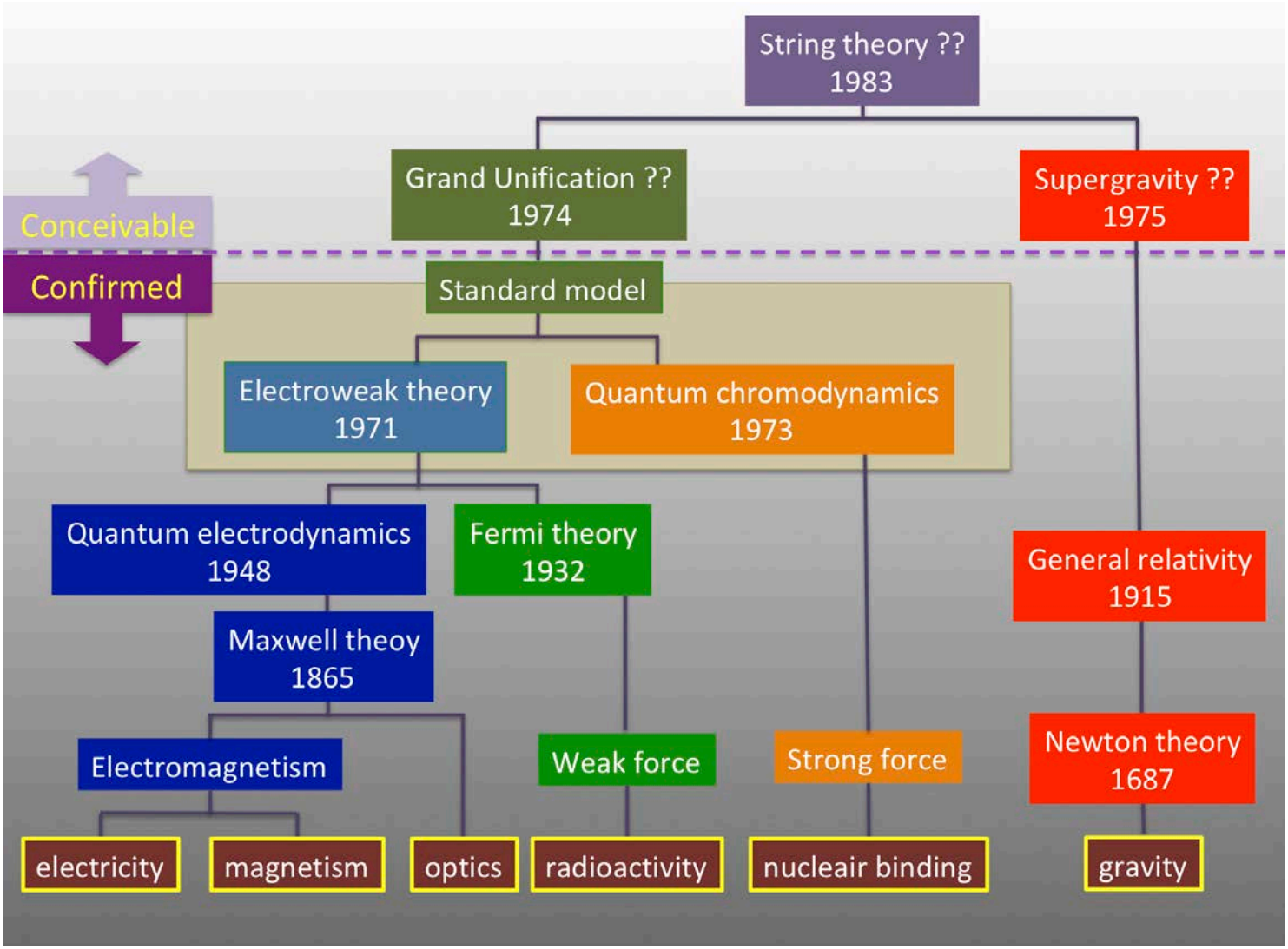


Figure I.4.45: *The well-established paths of unification that have led to the Standard Model, and conceivable paths beyond.*

Figure I.4.46: *Forces Unite!* A call for the United Forces of Nature.

dard Model but have not (yet) been vindicated by experiment and should therefore be labeled as speculative. The Standard Model has left us with a number of open questions that strongly hint in the direction of a single overarching quantum gauge theory that comprises the three non-gravitational interactions. Models of such type are called *Grand Unified Theories* or GUTs, but the proposals made so far have not been very successful. An example of such a hint is that the electric charges of proton and electron match perfectly, but within the Standard Model there is no a priori reason that they should have the same magnitude. Except when there would be magnetic monopoles around, but these are not part of the Standard Model; however in GUTs they exist. Another hint is that the family structure is not explained; it may possibly result from some underlying structure.

The theory of gravity remains a case apart. In spite of the tremendous successes of Einstein's theory, it has so far withstood all attempts to make it consistent with the principles of quantum theory. This is a highly non-trivial matter and seems to require a radical change of perspective. On the other hand it should not be too surprising: it is unique because it directly concerns the primary notions of space and time itself.

The line of development starting around 1970 centered around a few additional concepts: the first is the notion of *rigid supersymmetry*, the second was that of local or gauged supersymmetry which gave rise to *supergravity* theories, and finally the basic step from point particles to extended objects like *strings* and so-called *branes*. We close this chapter by a lightning review of some of the salient features of these developments.

Supersymmetry

From bosons to fermions and back. The gauge symmetries we have discussed so far transform certain particle types into each other. The $\mathrm{SU}(3)$ color group for example transforms the quarks of different colors into each other. The weak SU(2) transforms up and down quarks or electrons and their neutrinos into each other. But these gauge transformations always transform bosons into bosons and fermions into fermions. Supersymmetry is an intricate symmetry which involves generators which themselves are fermionic with the crucial property that they transform bosons into fermions and back. It entails a drastic extension of the notion of symmetry. Its discovery and early development goes back to the early 1970s. If we call the super charge (or generator of the supersymmetry) Q, it has the following properties:

$$\mathrm{Q}^2 = 0$$

$$\mathrm{Q}\,|\mathrm{boson}\rangle = |\mathrm{fermion}\rangle\ ;\ \mathrm{Q}\,|\mathrm{fermion}\rangle = |\mathrm{boson}\rangle\,.$$

One may add more supercharges, in which case we speak of *extended supersymmetries*. In four dimensions we have a maximum of $\mathcal{N} = 8$ supersymmetries. The more supersymmetry the more constrained the theory will be. Like with other symmetries particle types fall into representations of the various supersymmetry algebras and these

representations will contain both bosons and fermions. The smallest representation of $\mathcal{N} = 8$ extended supersymmetry contains a spin-two particle which is a natural candidate for the graviton. So in that sense there is a fundamental link between supersymmetry and gravitation.

The Dirac equation predicted the existence of *anti-matter*, and in a similar way supersymmetry predicts the existence of super mirror images of all the known particles. Bosonic particle species would have fermionic *superpartners* and *vice versa*. These partners are generally denoted as *sparticles*: *squarks* and *sleptons*, while the superpartners of the force particles are called *gauginos*, like the *photino*, the *Winos* etc. The corresponding fields are labeled by the same letters with a tilde on top and we have displayed some of them in Figure I.4.49. In that figure they are depicted as belonging to the massless sector of superstring theory that will be discussed shortly.

A Minimally Supersymmetric Standard Model (MSSM). One thing we can conclude immediately is that unfortunately the presently observed bosons and fermions (the inhabitants of the Standard Model listed in Figure I.4.35(a)) cannot be each other's superpartners, because the other properties do not match. It is like the situation with the Dirac equation where the proton could not be identified with the anti-electron because they have different masses. The proton (field) has its own Dirac equation. For a superpartner all intrinsic properties are the same except for the spin which differs by half a unit. So to make the world supersymmetric the very minimal thing one may do is construct the simplest $\mathcal{N} = 1$ supersymmetric extension of the Standard Model, and that means just doubling the panels of I.4.35(a) and put tildes on all the particle symbols. This Minimal Supersymmetric Standard Model (MSSM) is actively studied and a lot of effort is devoted to 'hiding' the unwanted partners and finding possible experimental signatures that show up in high-energy experiments. You see that, in particular with extended supersymmetries, one is forced to accommodate large numbers of new particle species. And to break the supersymmetry even more particles have to be added. We will refrain from discussing the MSSM in more detail.

The principal motivation to build the LHC at CERN was to find the Higgs particle, a crucial ingredient of the Standard model that lacked experimental vindication. But the physicists had another deep motivation and that was the hope that the LHC would allow for the much more revolutionary discovery of supersymmetry as an underlying principle of nature. So far there has been no evidence for this. If the 'sparticles' are really there, they would make up a shadow world, which is extremely weakly coupled to our discernible world. Not having seen them up to now means that the supersymmetry would have to be badly broken in our universe, because breaking can give a considerable mass to the super partners. It is a bit like the 'Higgs breaking' mechanism that gives mass to the W and Z particles that mediate the weak interactions in the Standard Model.

Yet, from another perspective it is not inconceivable that supersymmetry is a blessing in disguise. The lightest supersymmetric particle is absolutely stable by construction, and it has been suggested that this lightest supersymmetric particle, for example the *photino* (the super partner of the photon), is a candidate for the elusive particle that makes up dark matter. It couples very weakly, is neutral and massive, and makes a perfect *WIMP*, a *Weakly Interacting Massive Particle*, that is favored in many cold dark matter scenarios. We briefly discussed this in the section on cosmology in Chapter I.2 on page 76. What we may conclude at this point is that the discovery of superpartners in a lab like CERN or Fermilab would be a spectacular discovery in its own right, but would also put string theory (and supergravity) in a far more credible position as these theories predict their existence as a necessary ingredient of nature. We have to wait and see. One of the reasons science is demanding, is that it requires so much patience.

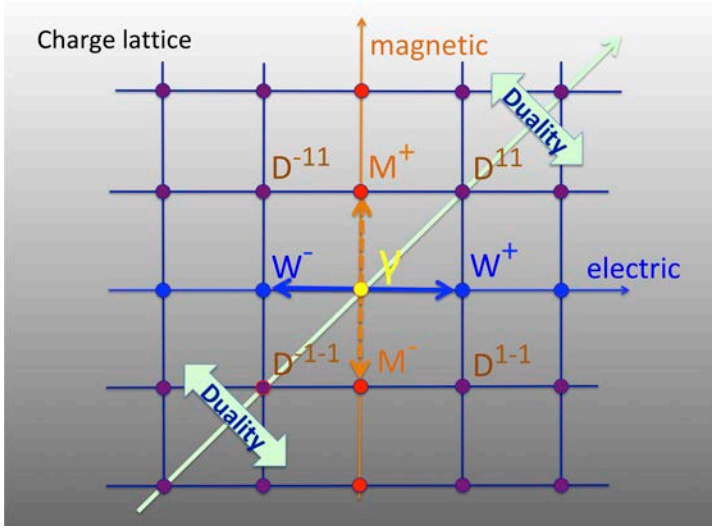


Figure I.4.47: *Montonen-Olive duality.* The electric-magnetic charge lattice of the $\mathcal{N}=4$ supersymmetric $\mathrm{SU}(2)$ Yang-Mills theory. This is in the weak coupling regime with on the horizontal axis a triplet of gauge bosons $W^{\pm}$ masses $\mathrm{M}_W = e\mathrm{f}$ and $W_0 = \gamma$ which remains massless. On the vertical axis there is the same massless photon and two magnetically charged monopole $\mathrm{M}^{\pm}$, these are solitons with mass $\mathrm{M}_{\mathrm{M}} = 4\pi\mathrm{f}/e$. The electric-magnetic Montonen-Olive duality corresponds to mirroring the lattice through the diagonal, interchanging the role of gauge particles and solitons and also interchanging $e \leftrightarrow 4\pi/e$; a strong-weak or S- duality.

To illustrate the power and beauty of supersymmetry, we briefly discuss two further examples: one is the $\mathcal{N}=4$ *supersymmetric Yang-Mills theory*, and the other is *supergravity* which plays a vital role in modern superstring theory.

$\mathcal{N}=4$ supersymmetric Yang-Mills. Let me briefly talk about a wonderful, somewhat exceptional class of models, which brings together a number of fundamental concepts that have been taking the stage in theoretical physics from the mid 1970s. The theories I am talking about are $\mathcal{N}=4$ *supersymmetric Yang-Mills theories.*

The marvel is that because of the $\mathcal{N}=4$ supersymmetry these theories are so constrained that their quantum behavior is well understood, even beyond the perturbative diagrammatic Feynman approach. This also implies that they exhibit an unusual kind of simplicity, which for the theorist makes them an ideal laboratory for testing novel ideas. It is for quantum field theorists what the roundworm C. Elegans is for geneticists so to speak. So it is not the theory on its own that is of particular relevance but its extraordinary properties are of interest

Let us consider the simplest case where the gauge group is $\mathrm{G}=\mathrm{SU}(2)$. The particle or field content of this $\mathcal{N}=4$ gauge theory consist of a single spin-one super-multiplet that transforms as a triplet or vector representation of the $\mathrm{SU}(2)$. Because of the supersymmetry, one can generate a *super-multiplet* by acting with the supersymmetry generators. The fields have the following spin content: there is one spin-1 field (these are the gauge bosons of the theory), there are four spin-$\frac{1}{2}$ and six spin-0, scalar and pseudo scalar fields. All of them transform in the triplet representation of the gauge group.

The scalar fields act like a kind of Higgs field and break the $\mathrm{SU}(2)$ gauge symmetry to $\mathrm{U}(1)$, which we call electromagnetism in analogy with the electro-weak theory. Because of the symmetry breaking two things happen:
(i) the gauge bosons $W^{\pm}$ acquire a mass $\mathrm{m}_W = e\mathrm{f}$, and $W^0 = \gamma$ is the massless $\mathrm{U}(1)$ 'photon'. The parameter f has dimension $[\mathrm{mass}]$ and sets the scale of the breaking. There is also a neutral massless scalar particle that survives in the breaking.
(ii) this theory has non-trivial classical *soliton* solutions, corresponding to the so-called '*t Hooft-Polyakov magnetic monopoles*. These are regular, finite energy classical field configurations that are stable for a topological reason, implying that magnetic charge is also strictly conserved. The monopoles $\mathrm{M}^{\pm}$ have a magnetic charge $\mathrm{g} = \pm 4\pi/e$ (twice the minimally allowed Dirac value) and have a mass (= energy of the classical field configuration) equal to $\mathrm{m}_{\mathrm{M}} = g\mathrm{f} = 4\pi\mathrm{f}/e$. Note that these magnetic monopoles are a

necessary ingredient of this theory, and you are not free to leave them out. They represent magnetically charged 'particle like' objects in the theory and what one may show is that upon quantization these monopoles also form spin-one supersymmetric representations.

Electromagnetic duality regained. What is striking in this theory is that on the quantum level it exhibits a dual symmetry between the electric and magnetic sectors of this theory. This is the non-abelian analog of the electric-magnetic duality of the source-free Maxwell equations that we mentioned in Chapter I.1, and is called the *Montonen-Olive duality.* We have depicted the duality transformation on the electric-magnetic charge lattice in Figure I.4.47, which shows the electrically charged gauge bosons $W^{\pm}$ and the charge neutral (self-dual) photon in the origin, as well as the magnetic monopoles $\mathrm{M}^{\pm}$ on the vertical axis. The spectrum also allows for dually charged sectors called *dyonic* labeled D(n,m). This remarkable symmetry is a *strong-weak* or so-called *S-duality*. Indeed if we take the electric coupling weak ($e < 1$), then the magnetic coupling is strong ($\mathrm{g} = 4\pi/e > 1$). So, this theory is like the pure Maxwell theory *self-dual*; it maps one to one onto itself under the duality transformation. The upshot is that we have two fully equivalent formulations of the same physics, one as the standard 'electric' gauge theory with massive $W^{\pm}$-bosons, a massless photon, and gauge coupling e, and the other as a 'magnetic' gauge theory with gauge bosons $\mathrm{M}^{\pm}$, a massless photon and a gauge coupling $\mathrm{g} \sim 1/e$.

Imagine what this means, if you turn up the coupling parameter e then you expect the strongly coupled theory to no longer be controllable and predictable. But in this case we have an alternative, not an alternative reality because there is only one reality, but an alternative perspective or description where that would-be violent and uncontrollable reality is very well behaved, completely calculable and predictable.

This special property derives from the fact that the theory is not only supersymmetric but also has *conformal symmetry.* This implies that the charges do not renormalize, and they do *not* develop a momentum dependence, like in the case of 'asymptotic freedom' of Figure I.4.40. The fact that the coupling constant has no dependence on momentum or distance means that this quantum theory is *scale invariant* ($\mathrm{r} \to \lambda \mathrm{r}$) and in fact *conformally invariant* because it is also invariant under inversion ($\mathrm{r} \to 1/\mathrm{r}$). It is a *superconformal gauge theory*.

There is one more point about this *superconformal gauge theory* which makes it even more exceptional. Remember that we mentioned that in addition to the massless photon we have also a massless scalar particle in the theory. This particle mediates an attractive force between the other particles with a coupling strength equal to the gauge coupling (the only coupling constant in the theory). Imagine we have two identical monopoles then we expect there to be a Coulomb repulsion due to the photon, but now there is the attractive scalar force which is exactly equal but opposite. And as you may have guessed, these two forces cancel each other out and that is truly remarkable. So, if you bring two monopoles together very slowly, they don't feel any force pushing them apart. The mass of a multiply charged monopole with charge mg scales exactly linearly: $\mathrm{M}_{\mathrm{mg}} = \mathrm{mM}_{\mathrm{g}}$. This implies that also the masses are not renormalized, and the classical mass formulas turn out to be exact. But adding a monopole with opposite magnetic charge is another story, because now the two forces add and the anti-pole feels an attractive force that is twice as strong. It is an unstable configuration, a monopole anti-monopole pair would annihilate and be converted into pure energy. And by the way for the charged particles like the W^{+} the same story holds.

So, that's the marvel: a supersymmetric , gauge and conformally invariant quantum field theory! A remarkable outlier, and indeed, some theorists feel tempted to quote Dirac's 1931 monopole paper, saying 'One would be surprised if

Nature wouldn't have made use of it.'

We will see that if not nature, then at least the string theorists have made use of it in a marvel called the AdS/CFT holographic correspondence that we will discuss later on.

Local supersymmetry: supergravity. A first important and profound generalization of Einstein's theory is a theory called *supergravity*, proposed in 1976 by Daniel Friedman, Sergio Ferrara and Peter van Nieuwenhuizen. Supergravity theories are invariant under *local* supersymmetry (or extended supersymmetry) transformations, and this means that the *supersymmetry is gauged*. It contains the Einstein theory but in addition to the *graviton* it predicts the existence of a fermionic partner called the *gravitino* with spin 3/2. These ideas have been worked out and extended in great detail ever since by a sizable community of devoted theoretical physicists. Extended supergravity theories would also encompass gauge symmetries of the Grand Unified type and were considered as candidates for a Theory of Everything. The maximally extended supergravity in four dimensions, which features 8 supercharges, is related to a unique supergravity theory in 11 dimensions, which was constructed by Eugène Cremmer, Bernard Julia and Joël Scherk working at the Ecole Normale Supérieure in Paris. The non-gauged $\mathcal{N} = 8$ theory in four dimensions can be obtained from the eleven dimensional one by compactifying seven dimensions on a seven-dimensional torus.

The sobering fact is that there was no support from the phenomenological side (no super symmetric partners ever showed up in experiments), and there is a myriad of extra particles that have to be accommodated (or better, eliminated) somehow. Moreover, the ultraviolet behavior of these theories of gravity kept causing problems. It turned out that they are not renormalizable, because of unwanted infinities that kept showing up in certain calculations. And this was resolved until much later, when around 1995 it was recognized that supergravity was the low energy approximation to a theory called *M-theory* living in eleven dimensions. This Meta theory, is the Mother of all ten--dimensional superstring theories which we will talk about shortly.

A Theory of Everything?

> Even if there is only one possible unified theory, it is just a set of rules and equations. What is it that breaths fire into the equations and makes a universe for them to describe? The usual approach of science of constructing a mathematical model cannot answer the questions of why there should be a universe for the model to describe. Why does the universe go to all the bother of existing?
>
> Stephen Hawking, *A Brief History of Time* (1988)

Let us recapitulate the big steps we have discussed in this chapter: we started with classical particles and classical fields like the electromagnetic field. Then we introduced the quantum theory, where we described basically a single particle in a fixed external force field and that produced an extremely successful model for the atom with electrons in orbits around the nucleus. Then we moved on to include the kinematics of special relativity and that brought us to quantum field theory where the distinction between force fields and particles was lifted, since both are described by quantum fields whose spectrum consists of states with an arbitrary number of particles of the type described by that field. This program culminated in the highly unified Standard Model. In their quest for an all-overarching *Theory of Everything (TOE)*, that would also include gravity, the physicists took one step further and started moving in various directions, all of which led up to the study of superstrings.

What if ....? The unification Figure I.4.45 at least suggests that a *Theory of Everything* is certainly not excluded. Hitherto a physical or logical veto that would prohibit such an overarching theory has not been disclosed. The term 'ev-

erything' is unfortunate, because it is not the pretension that such a theory would explain all observable physical phenomena, rather it would specify all the necessary and sufficient ingredients on a fundamental level, which would suffice to make a universe like ours. Still, you might wonder what it means, if such a Theory of Everything (or TOE) exists.

Most people think of it as a set of principles from which everything we do and do not know about nature would uniquely follow. We would take nothing for granted, not the existence of light nor certain particles, or even the existence of space and time. But maybe the starting point could be the notion of energy or information or of observability. As I said, *everything* is a lot, and practitioners aim a little lower. A TOE marks the end point of the quest for ever more fundamental and basic building blocks that are the subject of this chapter.

The discovery of a (or should I say *the*) TOE would mark the closure of basic physics. This would be both an impressive and a surprising achievement. Nature would have a true bottom so to speak. Yet, from a practical point of view such a completion is not such a big deal really. It probably would make physics a more boring place to be. Particle physics would at best become some kind of tourist trap, which one might want to avoid because the interesting characters lived there a long time ago. A monument for intelligence! And beauty, yes of course! Lots to admire and enjoy. But adventure? Alas, no!

But now I am talking like the physicists at the end of the nineteenth century who thought that the completion of basic physics was imminent. And it certainly was not! Quite the opposite, the twentieth century turned out to be one of the most revolutionary, inspiring and successful eras in physics ever. A century of relativity (geometry), of information and of quantum, as we argued in Chapter I.2.

Around 1980 it became clear that theories of 1-dimensional extended objects called *strings* provided a drastically different perspective on the problem of gravity. They have been center stage from 1984 onwards, but these theories have so far not been able to impress with resolving existing problems or with predictions that were confirmed by experiment. The relevance of string theory as a theoretical laboratory is fully recognized as a powerful extension of quantum field theory, and it has helped us to understand such elusive concepts as quantum black holes and quantum phase transitions. And superstring theory keeps alive the hope for a Theory of Everything, a Holy Grail of particle physics. Therefore we will conclude this chapter with a section on this topic which is still very much in a state of flux. As will become clear once more: beauty has its price.

Superstrings

> And so we face a contradiction between quantum field theory and general relativity similar to the contradictions that led to quantum mechanics. Many physicists believe that this contradiction contains the seeds of an upheaval as profound in its own way as the discovery of quantum mechanics and relativity
>
> Edward Witten, *Nature* (1996)

String theories in their present formulation are quantum theories of extended objects, like strings, and (mem)branes of different dimensions. Mathematical consistency of the this theory requires two conditions to be fulfilled, (i) the theory should be supersymmetric, and (ii) the theory lives in ten or eleven dimensions. A closed formulation of the theory that may exist in eleven dimensions, and for some mysterious reason is called M-theory, is not available, but a small set of ten-dimensional limiting descriptions of that theory are known, and these correspond to the five different superstring theories.

Figure I.4.48: *String worlds.* What the world might look like at $\simeq 10^{-35}$ m .

Just like a single quantum field describes an arbitrary number of particles of a given type, the basic property of the fundamental (super)string is that it describes an infinite number of fields, or particle types! Most of them correspond to extremely massive particles that cannot be produced in our accelerators. Crucial for phenomenology and falsifiability are the 'massless' fields that the theory predicts.
The one outstanding fact is that the theory includes gravity. The gravitational field obeys an equation to which Einstein's equations are an approximation. This makes the theory a serious candidate for a quantum theory of all fundamental particles and interactions including gravity. An important – quantessential – step forward, but many hurdles still have to be overcome. Most importantly, it is still not known how the beloved Standard Model fits in, though all the ingredients appear to be there. The problem so far is that the theory describes more than we need.

Understanding gravity. Our understanding and interpretation of gravity has through history made dramatic turns. Of course it started with the idea of a force leading to the whole Newtonian dynamical framework including his 'universal law of gravitation'. The second grand turning point came with Einstein's theory of General Relativity, where it was shown that the gravitational force was just a manifestation of the curvature of space-time. Further searches were driven by the strongly perceived necessity to bridge the gap between quantum theory and general relativity.

This turned into the elaborate field-theoretical edifice of supergravity in all its diversity. That approach turned out to have serious shortcomings and at some point seemed doomed, but then it gave way to superstring theory in which supergravity again found a safe haven.

String theory, as a possible overarching quantum theory of all interactions including gravity, has passed through some major revolutions after its inception dating from the early 1970s. It is customary to distinguish three eras of superstring theory:

1st era (...–> 1984): String theory as an attempt to describe the strong interactions (Veneziano, ...).
2nd era (1984-1995): Superstring theory as a theory of quantum gravity (Scherk et al, Schwartz, Green, Witten, ...).
3rd era (1995-present): Extended objects or D-branes, M-theory and the holographic AdS/CFT correspondence (Polchinski, Witten, 't Hooft, Susskind, Maldacena, Strominger, Vafa,...).

We see that string theory, as a would-be unified theory of all fundamental interactions including gravity, was launched in 1984. In that theory the gravitational field equations are derived from imposing conformal invariance on the underlying string degrees of freedom that live on the world-sheet of the string. In that perspective gravity is an effective long-distance description of an underlying string dynamics and in that sense is an emergent phenomena.

However, in spite of its intrinsic beauty and elegance, string

theory did not quite deliver. It forced us to accept many hard to swallow extras like supersymmetry, a 10-dimensional space-time, and quite some extra degrees of freedom, of which no hint showed up in experiments. A deluge of extra degrees of freedom that 'nobody had ordered' so to speak. But moreover it appeared that string theory had no direct answers in store concerning very important questions like how to treat *realistic* black holes quantum mechanically, and how to address the even more urgent questions concerning the direct experimental evidence for dark matter and dark energy.

And that takes us to the present engagement of string theory, in particular with the idea of holography which culminated in Juan Maldecena's rather stunning *Anti-de-Sitter/-Conformal-Field-Theory (AdS/CFT) correspondence.* This radical proposal was published in 1997 in a paper which is considered one of the most influential of the present era. It is often referred to as the *gauge/gravity duality* or *Maldacena duality*.

The gauge/gravity duality refers to a rather specific setting of the Anti de Sitter space-time (in various dimensions), but suggests a profound and generic aspect of string theory. The canonical example refers to the situation of string theory in the 5-dimensional *Anti de Sitter (AdS) space-time.* This space-time has a cosmic boundary, which is a flat 4-dimensional Minkowski space-time. On that boundary lives a four-dimensional conformal quantum field theory (CFT), which is a large N copy of the $\mathcal{N} = 4$ $\mathrm{SU}(\mathrm{N})$ gauge theory that we discussed in the previous section. The duality says that the full string theory in the AdS background is exactly dual to the CFT on the boundary. Thus the theories are fully equivalent; they describe the same physical reality in two different perspectives. So for example if we have the formation and subsequent evaporation of a black hole in the Anti de Sitter universe, this process could be completely understood as some unitary time evolution in that boundary conformal quantum field theory.

If you want, you can read the AdS/CFT correspondence in an even more – literally – 'outlandish' way, namely, that gravity and space-time are elevated to a holographic illusion! If we know everything about the conformal theory on the d-dimensional boundary, we would be able to reconstruct all conceivable (gravitational) physics in the (d+1)-dimensional space. This prompts the interpretation that gravity as such doesn't really exist as a fundamental force. How elusive can reality be? If it doesn't really exist, then it certainly wouldn't have to be quantized. The quantum behaviour is emulated in a quantum field theory living on the boundary of space-time. Let me paraphrase this ironic state of the universe as an ironic state of mind: or *we* are an illusion, or the theory that claims that we are an illusion is an illusion.

Strings: all fields in one?

What is a string? Let us start with the most elementary type of string which directly connects with our intuition. A string is like an idealized one-dimensional tiny piece of a rubber band that moves through ordinary space and time. The motion of a string can be broken down into the motion of its center of mass, and a relative internal motion. For *closed strings* the relative motion corresponds to waves moving in either direction *along* the string. But you can also have *open strings* that have to satisfy certain boundary conditions, which basically say that its endpoints have to move with the velocity of light or that they have to be attached to some higher dimensional physical object called a D-(mem)brane. These boundary conditions ensure that the string has a certain tension which is an energy per unit length. This tension makes these strings very much like the strings on a violin that have oscillatory modes, known as standing waves that correspond to its basic, harmonic overtones.

It is not hard to imagine how a string model is supposed to

represent particles, if we are far away – we cannot resolve the internal structure of the string, and we see the string as a point-like object with a certain mass and momentum and there may be other internal quantum numbers like charge or spin. Therefore strings manifests themselves as particles at large scales and low energies. The apparent mass of that particle corresponds to the energy of the internal (relative) oscillations of the string.

$$E^2 = (P^{com})^2 + \Sigma_k (p_k^{rel})^2 \simeq P^2 + \Sigma_k m_k^2 \,. \qquad \text{(I.4.38)}$$

Clearly, the different oscillatory modes (labeled by k) have to correspond to different particle species, but the mass scale of the masses m_k would be humongous. If the string is tiny, say of the order of the Planck length, then its internal modes are extremely hard to excite. You need an energy of the order of the Planck mass which we introduced in Chapter I.3, i.e. $m_k \simeq 10^{19}$ proton masses! We have to conclude that all the particles we know and love should correspond to different modes in the lowest energy or *massless sector* of the string. It is here that the superstring is important because it has a huge internal symmetry group which means that the massless sector is also extended, containing all spin values starting at two (the graviton) all the way down to zero. The zero mass sector of superstrings corresponds to the particle content of certain supergravity theories.

The take home message at this point is merely that a single string carries an infinite number of different particle degrees of freedom, of which only the massless sector is of phenomenological importance. So, one type of superstring may represent all different particle types and their superpartners as we have indicated in Figure I.4.49. So you should think of all the fields related to the particle types we have been discussing previously, corresponding to different modes of a single type of superstring. The higher mass modes are crucial to ensure that the theory is mathematically consistent, they help in making the theory well behaved at high energies. And that makes sensible calculations on the quantum level possible.

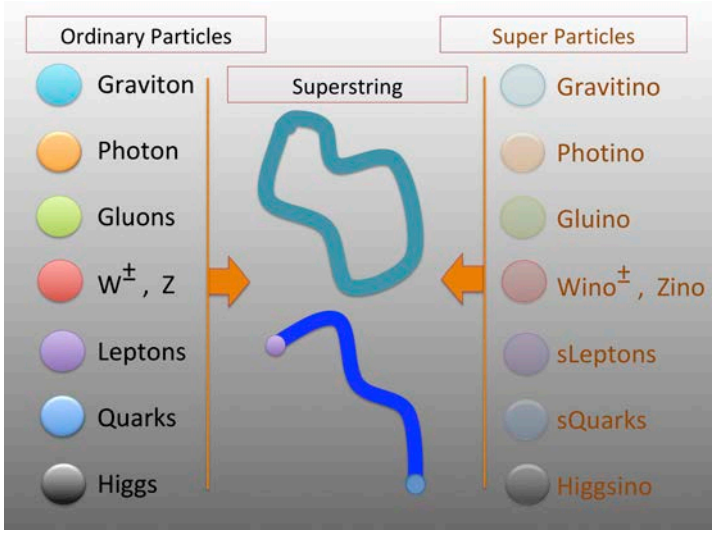


Figure I.4.49: *Superstrings.* All known particle types plus many more such as the superpartners or *sparticles* should correspond to different lowest energy modes of a superstring. These particle types were already ingredients of the earlier supergravity theories.

.

The world-sheet. If a point particle moves through space-time, we call its trajectory a *world-line*. Similarly if a string moves along in space-time, it traces out a two-dimensional surface embedded in space-time, a surface which is called a *world-sheet*. The world-sheet has one space-like dimension along the string and one time-like dimension to allow the propagation of the string.

There are two related geometries in the formulation of string theory: one is the two-dimensional intrinsic geometry of the world-sheet and the other is the geometry of the background space-time also called the target-space in which the string is moving. The world-sheet is parametrized by its space- and time-like coordinate (σ, τ) and its geometry is determined by a world-sheet metric $g_{\alpha\beta}(\sigma, \tau)$. This world-sheet is embedded in a ten-dimensional space-time with coordinates $(X^\mu; \mu = 0, ..., 9)$ with its own metric $g^{\mu\nu}(X^\lambda)$. The world-sheet is therefore described by its embedding,

$X^\mu = X^\mu(\sigma,\tau)$, that specifies its position in space and time. If you give me a point (σ_0,τ_0) on the world-sheet, the embedding yields a corresponding point $X^\mu_0 = X^\mu(\sigma_0,\tau_0)$ in space-time. A three-dimensional impression of what that looks like is given in Figure I.4.51. The embedding provides a relation between the two geometries in the sense that the embedding induces a metric on the world-sheet from the space-time metric. Just like we can construct the metric on the surface of a sphere by inducing it from the metric of the $\mathbb{R}^3$ in which we embedded the sphere, as we did in Chapter I.2.

The modeling of a string propagating in space-time involves an action (or Hamiltonian) that has all the required invariances and couplings, and loosely speaking it corresponds to the 'area' of world-sheet. This is not too surprising if you remember that the string has a tension and therefore wants to minimize its length and therefore energy (energy = length $\times$ tension).

What I want to get across here is that the expression for the 'area' of the world sheet involves the induced metric on the world-sheet which is an expression that in turn depends on the σ and τ derivatives of the space-time coordinates $X^\mu(\sigma,\tau)$. What this means is that in this formulation of string theory, the string dynamics is like a quantum field theory defined on the world-sheet, where the space time coórdinates X^μ play the role of a set of $(d+1)$ scalar fields. So, yes, we are indeed quantizing space-time in the sense that we quantize the coordinates. For superstrings the story is similar, a superstring moves in superspace which has also fermionic coordinates, and those provide fermionic field degrees of freedom on the world-sheet.

The string action has to be a scalar quantity and therefore will also involve the space-time metric $g_{\mu\nu}$, which depends on the space-time coordinates and makes the action highly nonlinear in the scalar fields. But let us for a moment assume we study the string in flat space-time then with $g_{\mu\nu} = \eta_{\mu\nu}$ is constant. Then the action will be invariant under space-time translations and Lorentz-transformations, and therefore we expect that the spectrum of the theory will reflect that and can be interpreted as representations of the Lorentz group and these label the space-time fields that the string theory produces. And that is for example how the graviton, as a massless spin-two representation, shows up in the spectrum of the string.

Background dependence. String theory goes fundamentally beyond General Relativity, because according to this theory space-time itself is supposedly made up of strings. In the actual formulation of superstring theory we have to deal with this paradox that on the one hand the strings propagate in a given background space-time, and on the other the actual background space-time is made up of strings. The background should be the outcome of the theory, it has to be predicted. This leads to certain consistency requirements. Space-time, as we experience it, is a manifestation of the collective behavior of strings, a kind of background or ground state. It would imply that space and maybe even time are 'emergent', an idea that would have tremendous philosophical implications as well.

In the massless sector of superstring theory we also find spin-one-half fermionic constituent particles, as well as the known spin-one force fields. Moreover these fields would couple in the correct way because the gauge symmetry principles that underly both the Standard Model and Einsteins gravity theory are naturally built into string theory. In a sense, string models have too much symmetry and therefore predict many extra particle types. We do not want these particles because we do not see them in nature; this implies that certain sectors of the theory have to be suppressed or even removed. To do that in a consistent way is a challenge.

As I have mentioned, for string theory to make sense two strong theoretical constraints have to be met. Number one is the existence of supersymmetry and number two a strin-

gent condition on the dimensionality of space-time. For superstrings the space-time dimension is ten. Indeed these conditions seem quite unnatural and make you worry, because neither supersymmetry nor ten-dimensional space-time have been observed. Yet from a mathematical consistency point of view there is no doubt about the necessity of imposing them from the start.

String interactions. In field theory the basic interactions are represented by interaction vertices where three or more particle lines meet and there is a coupling constant associated with each vertex. As strings represent all particle types their interactions should somehow take care of all particle interactions. In Figure I.4.50 it is clear that for the closed strings there is only one type of interaction vertex which corresponds to the joining and splitting of two strings and henceforth there is only one string coupling constant, called λ. The external legs of the string diagram, which represent the incoming and outgoing particles, are taken to be in the desired particle modes. This is how the different modes get coupled together and the complicated bookkeeping of labels is in some sense implicitly done by simply drawing the corresponding string diagram. A higher order string diagram with a number of incoming and outgoing strings is depicted in Figure I.4.51 .

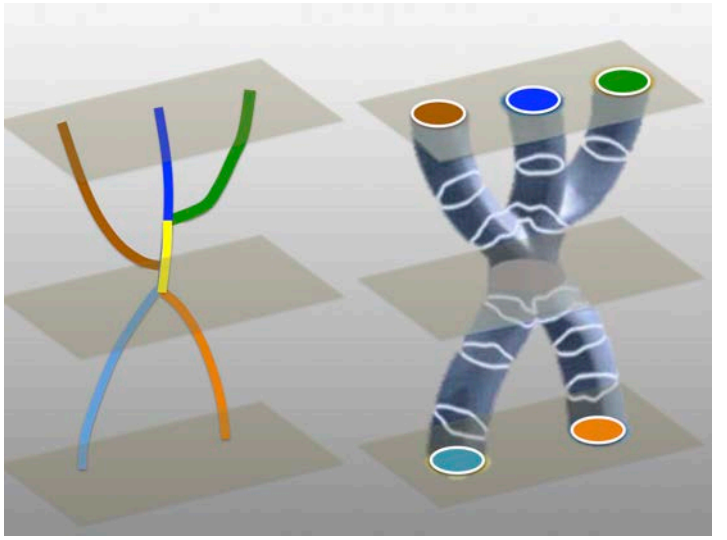


Figure I.4.50: *String interactions.* Strings have only one fundamental interaction vertex consisting of breaking or joining strings. So all different particle interaction diagrams of a given topology can be represented by a single string diagram.

.

String quantization.

Optimal paths. If you use Google maps it helps you find the shortest route. It offers you the choice between the shortest route in distance or the shortest route in time. What do particles do? If we take a photon, it will move from A to B along a path of minimal action, but what does that mean? The photon will certainly move along a straight line but that is the shortest in both time and distance. To really find out we have to do one more step. We know that in a medium like water or glass light moves slower than in vacuum or air. So, if A is under water and B above water the photons do *not* move along a straight line from A to B, they take a route that consists of two straight sections that make an angle at the surface. This is the problem we discussed in detail at the end of Chapter I.1 on page 18. The angle depends on the refraction indices in the two media, which are inverses of their velocities. The path that takes the shortest time is not the straight line from A to B, but rather a line that is broken at the surface. So the classical trajectories are optimal in the sense that they are minimal action trajectories, they correspond to local minima of the action.

That raises the interesting question that all school kids ask: How does a photon know which path to take? It can't do the necessary calculations, can it? No it can't! So it does not use Google's algorithm, which amounts to calculating most of the nearby paths in a restricted domain and choosing the optimal one. The photon, being a quantum particle does in fact a quantum computation it takes all paths simultaneously let them interfere and what comes out is weighted sum over possible paths the photon could have

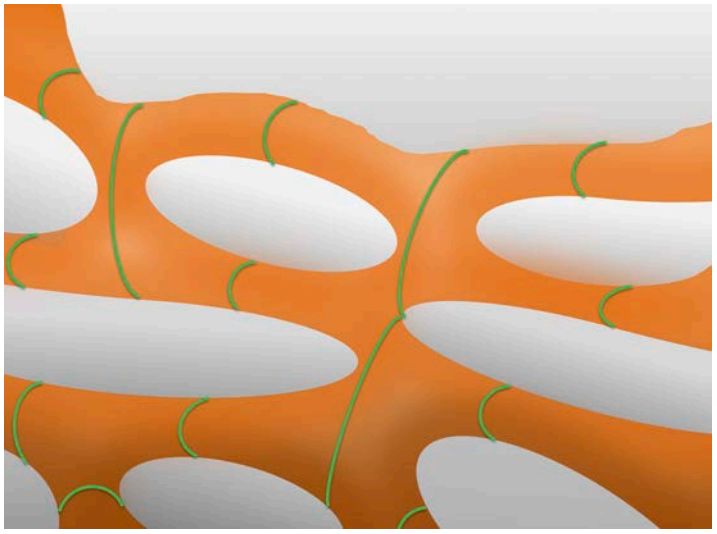


Figure I.4.51: *Joining and splitting strings.* Artist impression of strings moving in space-time and interacting by joining and splitting. It illustrates the geometric nature of string interactions. This string *world-sheet* has holes (handles) and closed boundaries which represent the incoming and outgoing closed strings.

taken. In short the photon performs a *path integral*.

The path integral. If we make the step to the quantum description of the particle, then we have to include all possible paths from A to B. The quantum propagation is now a weighted sum over paths, which is the famous Feynman path-integral. The probability amplitude to go from A to B is a superposition of amplitudes, the contribution of each path is weighted by an exponential phase factor where the phase is exactly the classical action of the path divided by $\hbar$.

So indeed in quantum mechanics these contributions can reinforce each other if they are in phase at B or dampen each other if they are out of phase; quantum particles interfere with themselves!

Going Euclidean. Here is another interesting aspect of path-integrals. In general they involve paths or configurations in space-time, which has a Lorentzian and not a Euclidean signature. However, what physicists often do when calculating or defining these integrals is to 'deform' them to Euclidean space, hence the term *Euclidean path-integral*. We calculate the Euclidean action of the paths and those with high action are exponentially suppressed with an exponent that equals *minus* the action divided by $\hbar$. One interesting consequence of this is that if we take the Euclidian action of a (d+1) dimensional physical system, we are summing over spatial configurations in a (d+1)-dimensional Euclidean space. But, as we argued in Chapter I.1 this is very much like doing statistical mechanics in (d+1) dimensions, where we for example calculated the partition function of the system as a sum of all possible configurations weighted by the Boltzmann factor which was also an exponential of minus the energy divided by kT. This correspondence expresses a profound relation between calculations in quantum field theory in d spatial and 1 time dimensions and calculations in statistical physics in (d+1) spatial dimensions. We will return to this connection in Chapter III.4.

From particles to strings. I tell you this particle story, because it helps to understand how string theory can be formulated as a generalization of a theory of point particles to a theory of one-dimensional extended objects. A classical path would typically correspond to a minimal action configuration of the world-sheet which corresponds to an extremal ('optimal') area of the word-sheet. The Euclidean equivalent for a closed string world sheet would be a soap bubble surface between two solid rings. And we know that a real soap bubble chooses the 'minimal energy' surface. It also showed what the fluctuations about the minimal energy configuration (the straight cylinder) look like: the string moves at intermediate times about its equilibrium position. There could also be wiggles running transversely – meaning along the string – but those have higher energy and unfortunately could not be excited with the soap-bubble kit I gave my daughter for her birthday in a failed attempt to make her study physics a long time ago.

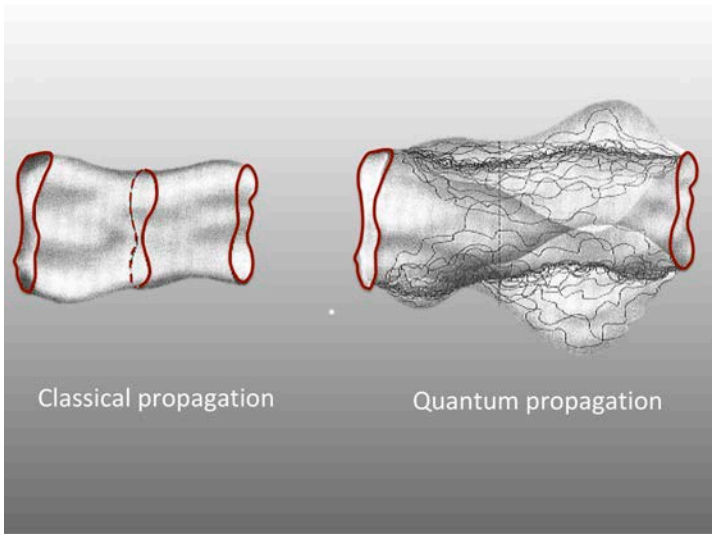


Figure I.4.52: *String propagator.* On the left the classical propagation of a string in space-time corresponding to the minimal action world-sheet bounded by the initial and final state. On the right the quantum propagation as a weighted sum over all string *world sheet configurations* bounded by the initial and final states. On the left there is a unique intermediate state, while on the right we have a superposition of many.

A string amplitude implies summing over all possible world-sheets satisfying the appropriate boundary conditions. This is illustrated in Figure I.4.52, where on the right we have a superposition of configurations that contribute to the propagation from the initial (left) to the final (right) configuration. The action is basically the area, which depends on the metric on the world-sheet. The problem then boils down to the construction of a correct and well-defined measure for doing this integral over the 'space of all metrics', without leaving metrics out, but at the same time not over-counting. This is a complicated mathematical problem because of the huge symmetries in the problem.

This basically concludes my ham-handed introduction to the quantization of strings including small world-sheet fluctuations.

Figure I.4.53: *Euclidean world-sheet.* Woman keeping up the Euclidean appearance of 'vacuum bubble,' a contribution to the vacuum-to-vacuum amplitude for a closed string. (Source: Atelier bulles geantes)

Weakly and strongly coupled strings. An important question at this point is, what are the world-sheet configurations that matter most, and will dominate the path-integral. In the figure just mentioned I have clearly limited myself to rather small fluctuations around minimal area classical configuration. So this is what a cheap navigator in your car would do, it misses out on surprising not so obvious shortcuts. You see for example that the topology of the world-sheets I included are all of the trivial cylindrical type. I have apparently not allowed for string interactions, meaning splitting and joining of tubes, and creating holes in the world sheet, like we depicted in Figure I.4.53 . What this basically means is that I have assumed that the string interactions are *weak*. The stronger the string interactions are, the easier (more probable) the excitation these complicated surfaces of high genus will be.

A full string amplitude requires summing over *all* possible world-sheets that satisfy the appropriate boundary conditions, which means that you end up with a sum over *genera*

(the number of holes) that label the topological class of the world sheet, and for any given genus you have to sum over all compatible metrics.

So as a relevant example let us consider the vacuum-to--vacuum amplitude for closed strings, This involves summing over all closed (no boundary) two-dimensional surfaces of arbitrary genus: these are called *Riemann surfaces*. They are embedded in (d+1)-dimensional Euclidean space and weighted by the negative exponential of their area. One such configuration with some holes features in Figure I.4.53 .[8] So this particular amplitude is now equivalent to an interesting problem in statistical physics: the calculation of the partition sum of random 2-dimensional surfaces in a d-dimensional space.

What these considerations make clear is that our naive intuitions about string theory really only apply to the case of weak string coupling, and nobody knows what a strongly coupled string theory actually means. That is to say, up to about 1995 nobody understood, but after the so-called second string revolution in the present era of string theory, we know much better what's going on. We will discuss some of this shortly.

Five superstring theories. You would maybe hope that with such outlandish requirements string theory would be highly unique, after all how could a Theory of Everything not be unique! But this appeared not to be the case. There are five different superstring theories in ten dimensions, which differ very much by the symmetries they have. Let us list these theories without further going into detail about their specific features: we distinguish: *Type I, Type IIA, Type IIB, Heterotic* $\mathrm{E}8 \times \mathrm{E}8$, and *Heterotic* $\mathrm{SO}(32)$. We have depicted these superstring theories and how they are connected to each other and with M-theory, to be discussed shortly, in Figure I.4.60 .

[8] A not-so-nice colleague suggested that this was part of a job application ceremony.

So, either these theories are wrong, or we have to work very hard to understand how our not-so-supersymmetric, not-so-ten-dimensional world can be interpreted as a not-so-simple solution to the equations that govern string theory. In that sense the theory does make very strong predictions, which at least in principle are falsifiable. This implies that we cannot just go out and do a decisive experiment, however, Predictions that have spent ages in waiting rooms are not uncommon in science, and this now also applies *a fortiori* to the very fundamentals of the quantum gravity world.

So far we have discussed (extended) supersymmetry as a distinguishing feature of supersymmetric gauge theories and super gravity theories. The next question is what the additional requirement that space-time be ten-dimensional means.

Ten-dimensional space-time? The second consistency condition of string theory is that space-time has to be 10-dimensional; this is in strong contrast to the 4-dimensional version we are all familiar with. At first this requirement sounds too outrageous to be true, and is all day convincingly falsified by our daily experience! But for theorists nothing appears as unsurmountable, they bear in mind Einstein's consoling words: 'Subtle is the Lord but malicious He is not.' And they have cooked up scenarios of how to get rid of six of those ten dimensions by a procedure called *compactification*. This amounts to tightly 'rolling up' the extra dimensions into a variety of compact manifolds like spheres or tori or combinations thereof.

Kaluza – Klein theory. This way of effectively reducing the dimension of space has a long history, and goes back to the first quarter of the twentieth century. Theodor Kaluza and Oskar Klein independently proposed a geometrical unification of the gravitational and electromagnetic forces, by looking at General Relativity in five dimensions, where they assumed that the fifth dimension would be curled up into a tiny circle. Interestingly, the extra components of

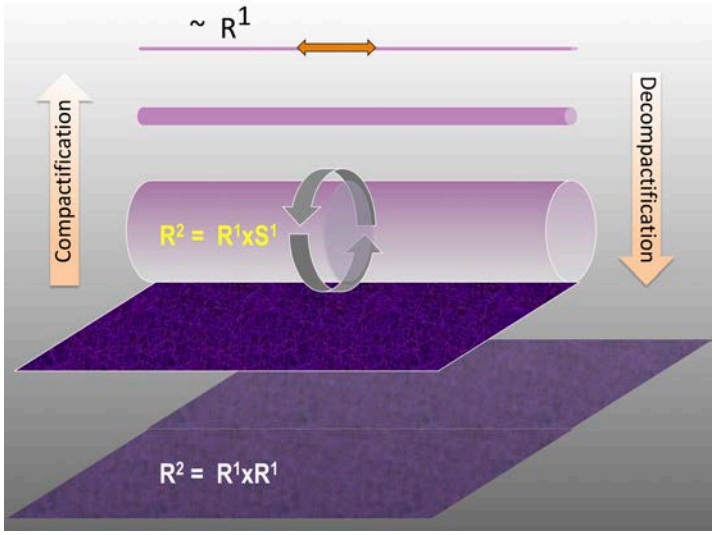


Figure I.4.54: *Compactification.* Here we give an idea how one spatial dimension can be compactified to a circle. The rotational U(1) symmetry of the circle results in the existence of U(1) gauge field in the large dimension. The relative coupling to the matter fields is inversely proportional to the radius R of the internal circle. In this perspective taking a weak coupling limit is like opening up an extra dimension ($R \to \infty$). This is the original dimensional compactification scheme of Kaluza and Klein, which plays also a role in going from 11-dimensional M-theory to 10-dimensional superstring theories.

the curvature tensor you get in going from four to five dimension would correspond to the degrees of freedom of the electromagnetic field in four dimensions (and an extra scalar field). The extra components of the metric would be '$g_{\mu 5} = g_{5\mu}$.' These generically correspond to the gauge potential A_μ, with g_{55} being an extra scalar field. Furthermore, the whole 5-dimensional system of Einstein's equations, after compactification, correctly reproduces the coupled Maxwell-Einstein equations in four dimensions. The momentum component in the fifth dimension of a moving particle basically corresponds to its electric charge. Einstein actually liked the idea but it was hard to reconcile with quantum theory and therefore slid into decline.

In science, attractive ideas that don't quite work can be safely stored away in a kind of fridge. This fridge consists foremost of the collective memory of the scientific community, and the written records of course. Ideas can hibernate for years or even for a century or so, before getting rediscovered and making a glamorous come-back in a novel context. Compactifying dimensions à la K-K is such an idea. The extra dimensions would probably be too tiny to see with present-day accelerators. To probe such small sizes you need correspondingly small wavelengths, which mean very high energies. So in that way you escape the manifest presence of those dimensions. But there is one feature that would be manifest at low energies. If the compactified space has symmetries – and it usually has – those symmetries after quantization give rise to massless particles that would clearly manifest themselves, also at low energies. It is precisely the rotational symmetry of the circle geometry of the fifth dimension that generates the massless photon in the Kaluza-Klein scenario!

Suppose we compactify one dimension of space into a circle then that has important consequences for the allowed states of a quantum particle. Remember that the spectrum of the momentum or energies for a free particle in flat space is continuous, and the wavefunction for a fixed momentum (p) state corresponds to a sinusoidal wave with the wavelength $\lambda = \hbar/p$. With the compactified dimension being a circle the particle momenta are quantized exactly as in the old Bohr model we discussed in Chapter I.3 on page 134, because we have the periodicity condition that the wave has to fit on the circle: $n\lambda = 2\pi R$ and therefore (relativistic) $E_n = p_n \sim n/R$. And in the original K-K model this component of the momentum is just the charge of the particle, that charge is thus quantized. The attentive reader now will ask whether here we have a model without monopoles where charge would be quantized. How come? The answer is that in this model topologically stable monopole configurations do exist as solitons much like in the supersymmetric gauge theory we discussed before. So Dirac does not have to turn in his grave!

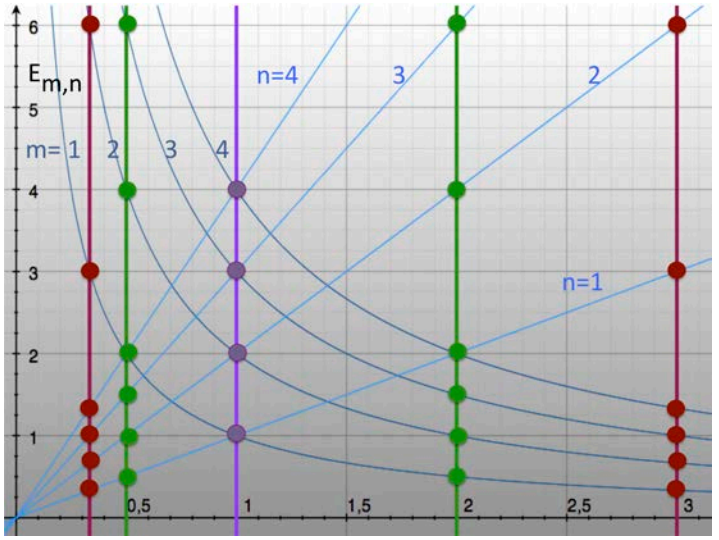


Figure I.4.55: *T-duality.* The non-oscillating modes of a of (bosonic) closed string on a circle with compactification radius R. The energy $E = E_m + E_n \simeq mR + \frac{n}{R}$, where m is the (topological) winding number and n the momentum which is also quantized. We have drawn the vertical lines for fixed R, demonstrating that the energy spectrum for radius R and $1/R$ are identical if we interchange m and n, making the spectra identical. Taking the limit $R \to \infty$ is decompactifying, which is like opening up an extra dimension. This means that the topological sector disappears and the momentum becomes continuous.

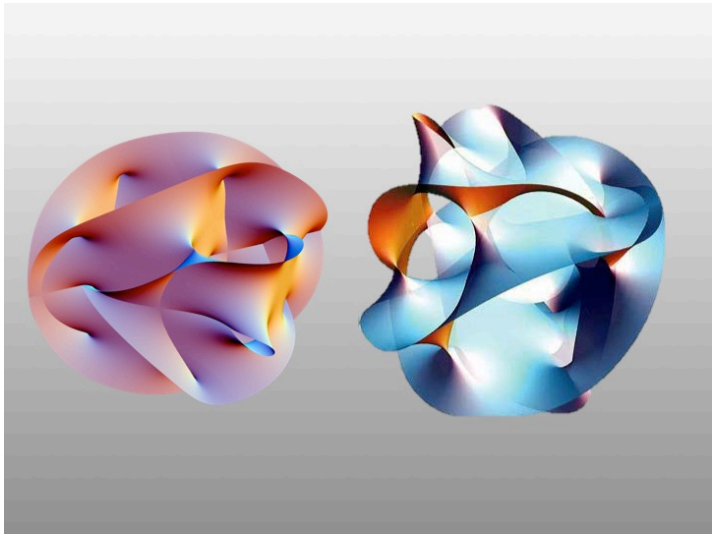


Figure I.4.56: *Compactification, the story of six inner dimensions.* Compactification means that space-time has four large and six compact or internal dimensions. Here we show two three-dimensional projections of possible six-dimensional compactifying spaces. It is evident that these so-called Calabi-Yau spaces have intricate geometries. (Source: Polytope24)

T-duality. If we have a string theory and we compactify one dimension, something interesting happens with the states of a simple string in that dimension. Of course the string can oscillate, but we are not interested in those states at this point. We want to look at the zero-modes. The string can move around the circle and behave like a particle, and that gives the spectrum we just discussed $E_n \sim p \sim n/R$, but for a string there are distinct topological sectors as the (closed) string can wind an integer number m times around the circle with radius R. This gives a contribution to the energy of the string proportional to its length so that gives a topological contribution $E_m \sim mR$. So, for a string we arrive (choosing appropriate units) at a simple formula for the energy spectrum of the non-oscillatory modes: $E = E_n + E_m = n/R + mR$. This spectrum is remarkable because it has an exact symmetry under the inversion $R \to 1/R$, as is shown graphically in Figure I.4.55.

And as the circle is part of the space-time, the target space in which the string moves, this symmetry or map is called '*T-duality*' or *target space duality*. Here we showed the elementary example where the duality was actually a symmetry, a self-duality, but the duality as a map plays an important role in the mapping of different 10-dimensional superstring theories onto each other as we will see shortly.

Note that we have encountered two types of duality: the first was called 'S-duality' or 'strong-weak duality' which may apply to both supersymmetric particle and string theories. Secondly we have 'T-duality' or 'target space duality', which makes only sense for string theories.

Calabi-Yau compactifications. An interesting quite special class of compact six-dimensional spaces over which the superstring can be compactified is the class of 6-dimensional Calabi-Yau manifolds, used in the compactification from ten to four space-time dimensions. These spaces have a number of very special properties that we will not discuss here. In Figure I.4.56 we exhibit two-dimensional views of three real dimensional cross-sections of these six-dimensional spaces. They are of interest for certain superstring theories, because they ensure that the resulting four-dimensional theory closely resembles a super extension of the Standard Model.

The multiverse. A weakness of the compactification scenarios in string theory is that nobody has been able to show that compactifications – if any – would be generated dynamically by this prospective Theory of Everything. This is a pressing issue because what looked like a unique and universal theory turns out to have an astronomical proliferation of conceivable compactifications. But to each compactification would correspond a different type of four-dimensional universe, with its own cosmic history, particle content and set of forces, in short, its own Not-So-Standard Model! Some of these universes could collapse before anything interesting would happen. Others may expand too fast for stars to form, let alone life to develop. In some of them there would be light, while in others none, or maybe many sorts of light. In fact a mind-boggling universe of universes is opening up, which is called the *multiverse*.

String theory so far is a theory of many possible theories, a theory of a multiverse in which wildly differing types of physics could manifest themselves, even in parallel. And, yes, ours would be just one of them. This is quite orthogonal to the basic motivation of most scientists who search for a unique universal theory of Nature. The common prejudice used to be that the Theory of Everything would come up with our dear universe as the unique or at least strongly favored solution. We like to think of our world as the unique

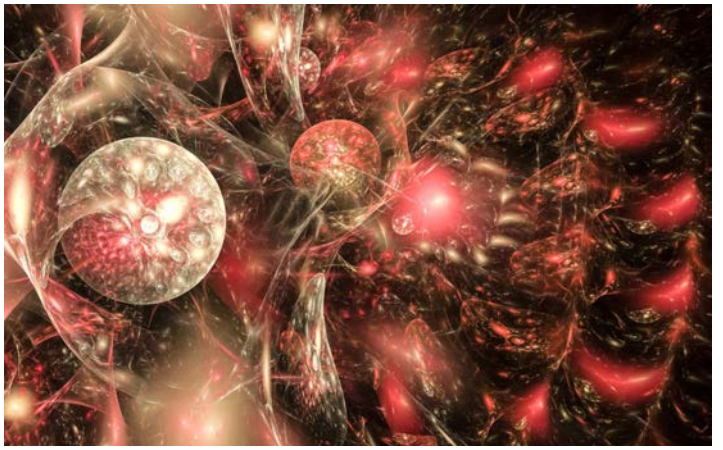


Figure I.4.57: *A multiverse?* An artists impression of the multiverse. A two-dimensional projection of a ten-dimensional compactification scheme. (Source: Forum Futura)

expression of the universal principles underlying that theory. It would have left the Creator but one option. It came somewhat as a shock that, after a promising start in that all-overarching direction, string theory moved in fact the opposite way. Maybe it is trying to tell us something contrary to our expectations, and for that science has excellent credentials. After all, the existence of a multiverse is yet another step away from our old anthropocentric dream of us being (in) the center of THE universe. THE universe? What are you talking about? Such a dramatic form of relativism would indeed constitute the ultimate irony of science, or of human existence.

It is not by accident that a strong protagonist of the multiverse, Leonard Susskind of Stanford University who wrote a popular book called *The Multiverse* about it, claims that the proper interpretation and *main prediction* of string theory is precisely that we live in a multiverse. The nasty aspect of this view is that the existence and properties of our own universe become extremely hard to predict from such a premise and in that sense not much progress has been made. It is an example of where contingency and evolutionary thinking enters physics at the most fundamental

level. It is analogous but much worse than asking a physicist to precisely predict the number and sizes of the planetary orbits in our solar system from Newton's laws. That can't be done. Newton's equations describe all sorts of planetary systems. And indeed, all observed systems do fit in his framework, but asking to predict them would be many bridges too far. And to be fair we don't expect that because we know that the details of the solar system are the outcome of a highly contingent, non-universal historical process and not simply calculable from first principles. To put it differently, all dogs are animals but not all animals are dogs, that's the problem! The biological conundrum, we know what an animal is and what the species on Earth look like, but predicting them from single-cell organisms using genetics is somewhat harder.

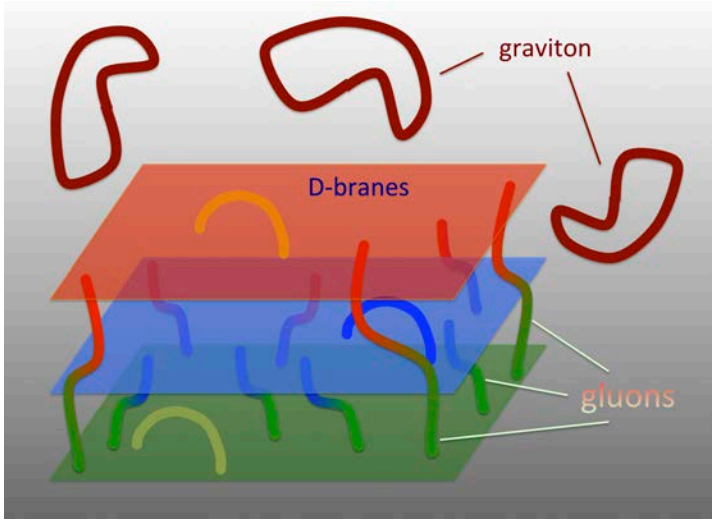


Figure I.4.58: *D-branes and strings.* A stack of three flat D2-branes. Open strings have to end on branes, by connecting them they represent nine $\mathrm{U}(3)$ gauge fields living on the branes. This figure is reminiscent of Figure I.4.37 with the bi-colored lines representing gluons propagating. Closed strings are not connected to branes and correspond to gravitons.

M-theory, D-branes and dualities

> It was believed for many years that there were five possible string theories, prompting the question: if one of these describes our universe, who lives in the other four worlds? But recently it has become clear that those five string theories are limiting cases of one majestic and mysterious theory.
>
> Edward Witten, *Nature* (1996)

D-branes. Whether they liked it or not, string theorists discovered that strings are not enough to make a consistent quantum theory of gravity. In fact 11-dimensional supergravity had a somewhat uncomfortable 'living apart together' relation with the 10-dimensional string theories. This supposedly low energy approximation of string theory lived one dimension up and had features that were lacking in string theory. These were stable soliton like classical solutions called branes, to be thought of as p-dimensional generalizations of membranes. This quite naturally prompted the question what the role of these *p-branes* in string theory would be.

Strings are one-dimensional extended objects, and indeed, a question that came up already early-on was: why if you give up the unique particle notion as the fundamental starting point, stop at one-dimensional extended objects? Why not include membranes and other higher dimensional extended objects? It was part of the second string revolution around 1995 that Joe Polchinski of the Kavli Institute for Theoretical Physics in Santa Barbara had the crucial insight that higher dimensional objects he called *D-branes* had to be included indeed. He intoduced D-branes in string theory as the end points of open strings. They could therefore in principle have dimensions p running from zero to nine. D-branes could be flat of infinite extent, or curled up into compact objects like black holes for example, they could be single or stacked up. Each type of string theory would allow for D-branes of specific dimensions. In superstring theory these p-dimensional D-branes, or Dp-branes, are dynamical objects which in the appropriate su-

pergravity limit correspond with the very heavy soliton-like p-banes.

In Figure I.4.58 we have depicted a stack of three colored D2-branes, and we have also drawn open strings ending on them. Let us discuss this picture in the weak coupling low energy limit of string theory, then it is known that open strings carry a vector representation of supersymmetry algebra. These open string states correspond to fields carrying one space-time index μ in this case running from 0 to $p = 2$, and carry two 'internal' (color) indices labeling the branes to which they are connected, in other words they beg to be identified as gauge fields A_μ^{ab}. In the figure there are nine possible color combinations, making up a three-dimensional $U(3)$ 'color' gauge theory. The picture clearly shows that the gauge theory is attached to the Dp-brane and therefore (p+1)-dimensional. From the branes point of view the strings between them are like excitations of the branes, they describe the brane dynamics. If the D-brane is embedded in a higher dimensional space then we have that the closed strings live in a higher-dimensional $(d > p)$ space then the gauge fields as indicated in the figure. This fact posed yet another conundrum one had to face. One additional comment on the figure, imagine the the D-branes to be so-called *black branes* meaning that they would correspond to some horizon then one could imagine the open strings pairing up to make some closed strings which could then leave the D-brane. The brane would radiate!

M-theory and superstring dualities.

> This theory, which is sometimes called M-theory (according to taste, M stands for magic, mystery, marvel, membrane or matrix), is seen by many as a likely candidate for a complete description of nature.
>
> Edward Witten, *Nature* (1996)

In Figure I.4.60 we give a bird-eye's view of the modeling landscape in ten and eleven dimensions. Who is living where and how they are related. The precursor of string theory was 11-dimensional supergravity: it did have attractive features like for example classical 2- and 5-brane solutions but seemed to not be fully consistent. In the figure it has moved a bit to the background because it is presently understood to be the low energy approximation of M-theory. The magic theory that remains in many ways a mystery, as there is no explicit formulation available, and we don't even know if such a formulation exists. And therefore the most fundamental principle underlying M-theory may still be hidden as well.

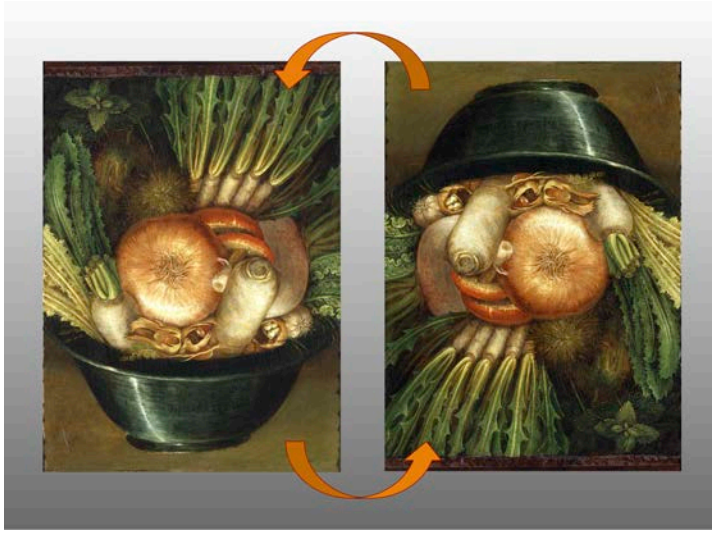


Figure I.4.59: *Duality.* In this painting of the Renaissance Milanese painter Guiseppe Arcimbolo (1526 – 1593), we see that a given physical vegetarian reality, this particular painting called *Verdure* or *Vegetables*, has two different interpretations which are dual to each other. In the weak coupling limit it is a vegetable basket, in the – I presume – strong coupling limit it turns into a vegetable face. The transformation is a rotation over an angle of π. (Source: ©Photo Scala, Florence.)

M-theory is known because of its low energy supergravity manifestation and through other limits which manifest themselves as the five superstring theories in ten dimen-

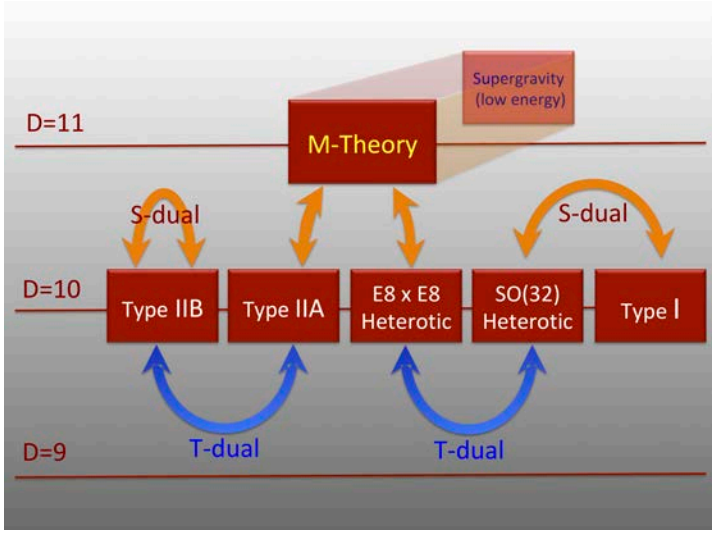


Figure I.4.60: *M-theory.and string dualities.* M-theory is a theory in eleven dimensions. It contains all five 10-dimensional superstring theory types. These ten-dimensional theories can be related through certain S- and T-dualities.

sions. These limits are related to compactification of one space dimension. For example the low energy limit – supergravity – can be compactified over a circle to ten dimensions, where the supergravity 2-brane is wrapped around that circle, which reduces the 2-brane to a string. And the resulting theory could be identified as the weakly coupled ten-dimensional *Type 2A* string theory. In the strong coupling limit of the string theory the compactified dimension would open up as we have discussed before. Furthermore, in the *Type IIA* string theory one could do a T-duality transformation (like $\mathrm{R} \rightarrow 1/\mathrm{R}$), which turns the theory into the *Type IIB* string theory. So it is in this sense that many connections between the various models were established, and indeed this network of dualities clearly demonstrated that these five theories are basically five different guises of one underlying theory, which has been named M-theory. Quite a mind-boggler!

At this point of the tour we have arrived at the center of the third string era, and I could imagine that if you are a 'freshman' reader, not at all familiar with these ideas, this narrative will come across as an arcane, brilliant but bizarre endeavor. A type of excursion in domains of the mind that you would not expect in a book about physics, a discipline that stands out for its factual rigour and its exemplary strong empirical basis.

The reason that I include these developments is exactly because this is what the struggle of science at the frontiers of knowledge may look like and should look like. It should be explorative in all conceivable ways, as long as it is not plainly stupid. This holds for the experimental as well as theoretical domain. It was Einstein who in 1934 made the following remark:

> The theoretical scientist is compelled in an increasing degree to be guided by purely mathematical, formal considerations [...]. The theorist who undertakes such a labor should not be carped at as "fanciful"; on the contrary, he should be granted the right to give free rein to his fancy, for there is no other way to the goal.

Let us in this vein explore a final set of fancy ideas that got a spectacular impetus out of string theory.

Holography and the AdS/CFT program

> We would like to advocate here a somewhat extreme point of view. We suspect that there simply are no more degrees of freedom (inside a volume) to talk about than the ones on can draw on its surface as given by $\mathrm{S} = \mathrm{A}/4$. The situation can be compared with a hologram of a three-dimensional image on a two-dimensional surface. The details of the hologram on the surface are intricate and contain as much information as it is allowed by the finiteness of the wavelength of light–read the Planck length.
>
> Gerard 't Hooft *Salamfestschrift* (1993)

Holography. We are going to discuss a profound novel correspondence that is of interest to physics and mathematics of many sorts. It is the holographic idea, that all the information contained in a space of $(\mathrm{d}+1)$ dimensions can be faithfully represented on a holographic screen of d dimensions. In the context of gravity, this conjecture was first put forward in the quantum understanding of black holes by 't Hooft as early as 1993 and taken up shortly after by Susskind in the context of string theory. We have talked about black hole entropy and the information paradox in the section on black holes in the previous chapter on page 139. The current narrative is that information cannot get lost but instead is somehow encoded, 'frozen in', on the horizon. The horizon keeps track of all things that pass by, so to say. This information content corresponds exactly with the Bekenstein-Hawking entropy that is also located on the horizon of the black hole. It would allow for the possibility that the information would ultimately be carried away again as hidden correlations in the Hawking radiation. So the information carried by things that have fallen into a black hole can – in principle at least – be retrieved. In particular in a consistent quantum mechanical description of the black-hole formation and evaporation process this *has* to be the case.

Black hole holography. The study of holography applied to black holes in the context of string theory culminated in a 1997 paper by Juan Maldacena at the Institute for Advanced study in Princeton, in which he made a strong claim of an exact dual relation between a superstring theory in a 5-dimensional Anti de Sitter space denoted as AdS_5, and an $\mathcal{N} = 4$ supersymmetric $\mathrm{SU}(\mathrm{N})$ gauge theory for large N defined on the boundary of AdS_5, which is equivalent to four-dimensional flat Minkowski space. This exceptional claim has in the meantime been substantiated and extended in many very convincing ways.

It brings together a number of ideas that we have touched upon in this book: the idea of symmetric curved space-times that are solutions to Einstein's equations, the idea

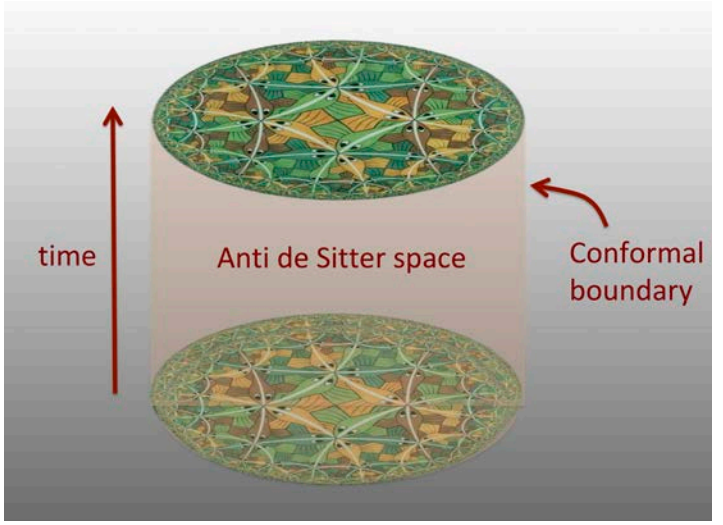


Figure I.4.61: *AdS/CFT correspondence.* The superstring theory compactifed over S^5 to a five-dimensional Anti de Sitter (AdS) space-time, which corresponds to the interior of the cylinder. That space has a 4-d boundary which is a flat Minkowski space (the cylindrical surface) on which a conformal field theory (CFT) lives, which is the hologram, a fancy encoded but faithful representation of the five-dimensional string theory in the interior.

of supersymmetry and supergravity, the $\mathbf{N} = 4$ superconformal gauge theories, and the ideas of string theory/supergravity compactification. Indeed a more encompassing confluence of ideas is hard to imagine and it may rekindle dark memories of some horrifying final exam that you once failed. I am sorry!

To be slightly more specific we are discussing the 10-dimensional IIB string and supergravity models, which are defined on a 10-dimensional space-time $\mathbf{M} = \mathrm{AdS}_5 \otimes S^5$, then on the AdS space one ends up not with 8 but 4 supersymmetries, the other four get broken by the compactification. Furthermore the background space AdS_5 is a very special space that has a large symmetry group which corresponds with the four-dimensional conformal group, and this yields a space-time theory with a $\mathbf{N} = 4$ superconformal symmetry.

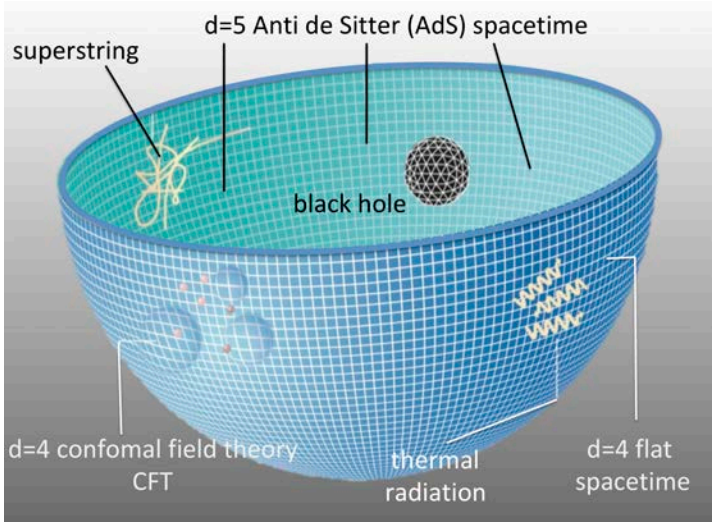


Figure I.4.62: *AdS/CFT correspondence.* An artist's impression of the holographic principle as it is operative in a compactification from ten to five dimensions. The superstring theory lives in a five-dimensional Anti de Sitter (AdS) space-time (inside the bowl). That space has a 4-d boundary which is a flat Minkowski space (the blue bowl) on which a conformal field theory (CFT) lives, which is the hologram, a fancy encoded but faithful representation of the five-dimensional string theory in the interior. (after A.T. Kamajan)

This string theory has actual open and closed strings but also D4-branes and we put a stack of N at the boundary, so producing an $\mathbf{N} = 4$ super-conformal SU(N) gauge theory at the boundary. Note that the symmetries of both theories coincide and therefore that will facilitate the identification of string states in the bulk with particular observables in the quantum field theory.

This gauge/gravity duality is a strong-weak type duality, which means that the low energy weakly coupled gravitational theory tells us about the strong coupling behavior of the quantum field theory. And as there is a complete equivalence the converse is also true, so the weakly coupled gauge theory should teach us about strongly interacting string theory.

If you look at the Hilbert space of states or the spectrum of the string theory on AdS_5, you get an impression how involved and surprising this Maldacena correspondence must be. For very low energy the theory of gravity just is the Einstein equations linearized around the background, and the excitations are gravitational waves. If we move to string theory, we get the closed strings which represent the massless supergravity degrees of freedom and if the background has D-branes we will excite open strings attached to branes. After that massive string modes will also be excited. When we go up even further in energy we will enter the regime where D-branes will be created. And if they are sufficiently heavy, they may start to form small black holes. And the more energy we put in the heavier and larger the black holes become. This process can continue until the horizon coincides with the boundary of a very large black hole. To imagine that this great variety of interacting degrees of freedom can be faithfully mapped to a large N super-conformal gauge theory is quite miraculous. What I can tell you is that a great variety of checks has been performed (involving very extensive computations of particular features) and all of these have confirmed the expectations.

However, the real shortcoming of this correspondence is that it only seems to work in this quite exceptional geometry with the serious problem that it involves the *Anti* de Sitter space which has a *negative* cosmological constant. This is in direct contradiction to the well-established fact that our universe has a small but definitely *positive* cosmological constant. This appears to pose a serious challenge to the AdS/CFT programme. The question is whether there is some version of a duality that holds also in De Sitter space.

Emergent gravity. This brings us to a brief description of the still rather controversial idea of 'emergent' or 'entropic gravity' and its protagonists like Erik Verlinde and collaborators. They have addressed the question of what could happen if we move from an Anti de Sitter to a De Sitter

background with positive cosmological constant. It is suggested that a rigorous holographic image on the boundary is no longer what happens, instead the speculation is that the entropy will acquire a volume term and spread from the boundary into the space, maybe causing deviations from Newtons laws that look like the effects of dark matter and dark energy. If confirmed by further analysis that would certainly constitute another truly stunning result.

At home in the quantum world

> Before I came here I was confused about this subject. Having listened to your lecture I am still confused. But on a higher level.
>
> *Enrico Fermi*

We have come to the end of the first part of the book. This part was devoted to the basic concepts and contents of quantum theory and its classical roots, as it developed over the last century. We of course always have lived in a quantessential world, but it is only now dawning upon us what that means. We started by describing the gems of classical physics, which ran into a number of serious troubles that could only be resolved by embracing the quantum principles. In this chapter we described the subsequent successes of applying the quantum principles to ever deeper layers of the microscopic world. A journey that as we saw is by no means completed.

Quantum theory entered our thinking on the atomic scale, say at 10^{-10} meters, and from there it started spreading. We recall that there are two ways to go from there and extend the results. The first is to go to ever smaller scales, and that is the route we have followed in this first part of the book. We went all the way down from the atom, via the nuclear structure to the elementary quarks and leptons, to a scale of about 10^{-20} meters, the scale accessible to modern accelerators like the LHC at CERN.

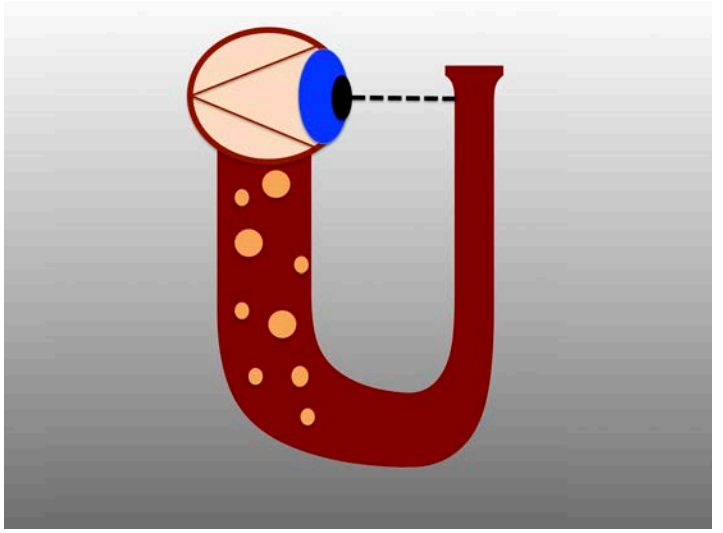


Figure I.4.63: *It fom Bit.* In a famous essay John Archibald Wheeler pondered over the philosophical ramifications of the idea that Information lies at the basis of our universe. Is it possible that All of Nature grew out of information only? The Cosmic Code as an all-embracing Hyper Genetics. This intriguing image also symbolizes evolution, or how Nature seems to be in search of itself, through the human effort of scientific inquiry, which by definition is part of that Nature.

The other way is to go up and apply quantum theory on scales corresponding to chemistry, or the many other forms of condensed matter that we find in nature or create in the lab. We save this part of the quantum story for the final Volume. The next, middle part of the book is called *quantessence*, and is devoted to the more formal aspects of quantum theory. We will expose some of its very rich logical and mathematical structure and comment on it. I think it would be a poor choice to leave it out, exactly because it is a central part of quantum theory. We can't really do without because what makes the theory so attractive is that on a conceptual level it is so counter-intuitive. And where confusion reigns it is vital to keep the language as clean as possible as to be sure about what the questions are and what the answers mean. I don't know of any theory where the mathematical framework is so rich and unambiguous,

and at the same time the narratives and interpretations are so paradoxical and hard to grasp. This makes the theory exciting and we will discuss a number of well-known but stupefying paradoxes in the next Volume. We are all set to start climbing the amazing mount quantum.

> If the world 'out there' is writhing like a barrel of eels, why do we detect a barrel of concrete when we look? To put the question differently, where is the boundary between the random uncertainty of the quantum world, where particles spring into and out of existence, and the orderly certainty of the classical world, where we live, see, and measure? This question...is as deep as any in modern physics. It drove the years-long debate between Bohr and Einstein. . . . Every physical quantity derives its ultimate significance from bits, binary yes-or-no indications, a conclusion which we epitomize in the phrase, *it from bit*.
>
> John Archibald Wheeler,
> *Geons, Black Holes & Quantum Foam* (1998)

Further reading.
On nuclear physics:

- *Introductory Nuclear Physics*
  Kenneth S. Krane
  Wiley (1988)

On particle physics:

- *Introduction to Elementary Particle Physics*
  Alessandro Bettini
  Cambridge University Press (2014)
- *Concepts of Elementary Particle Physics*
  Michael E. Peskin
  Oxford University Press (2019)

On string theory:

- *The Elegant Universe - Superstrings, Hidden Dimensions, and the Quest for the Ultimate Theory*
  Brian Greene
  W. W. Norton & Company (2010)
- *The Little Book of String Theory*
  Steven S. Gubser
  Princeton University Press (2010)

Complementary reading:

- *The Second Creation: Makers of the Revolution in Twentieth-Century Physics*
  Robert P. Crease
  Page (1991)

Indices

Subject index Volume I

Name index Volume I

SANDER BAIS

Quantessence: how quantum theory works

The Quantessence of Reality

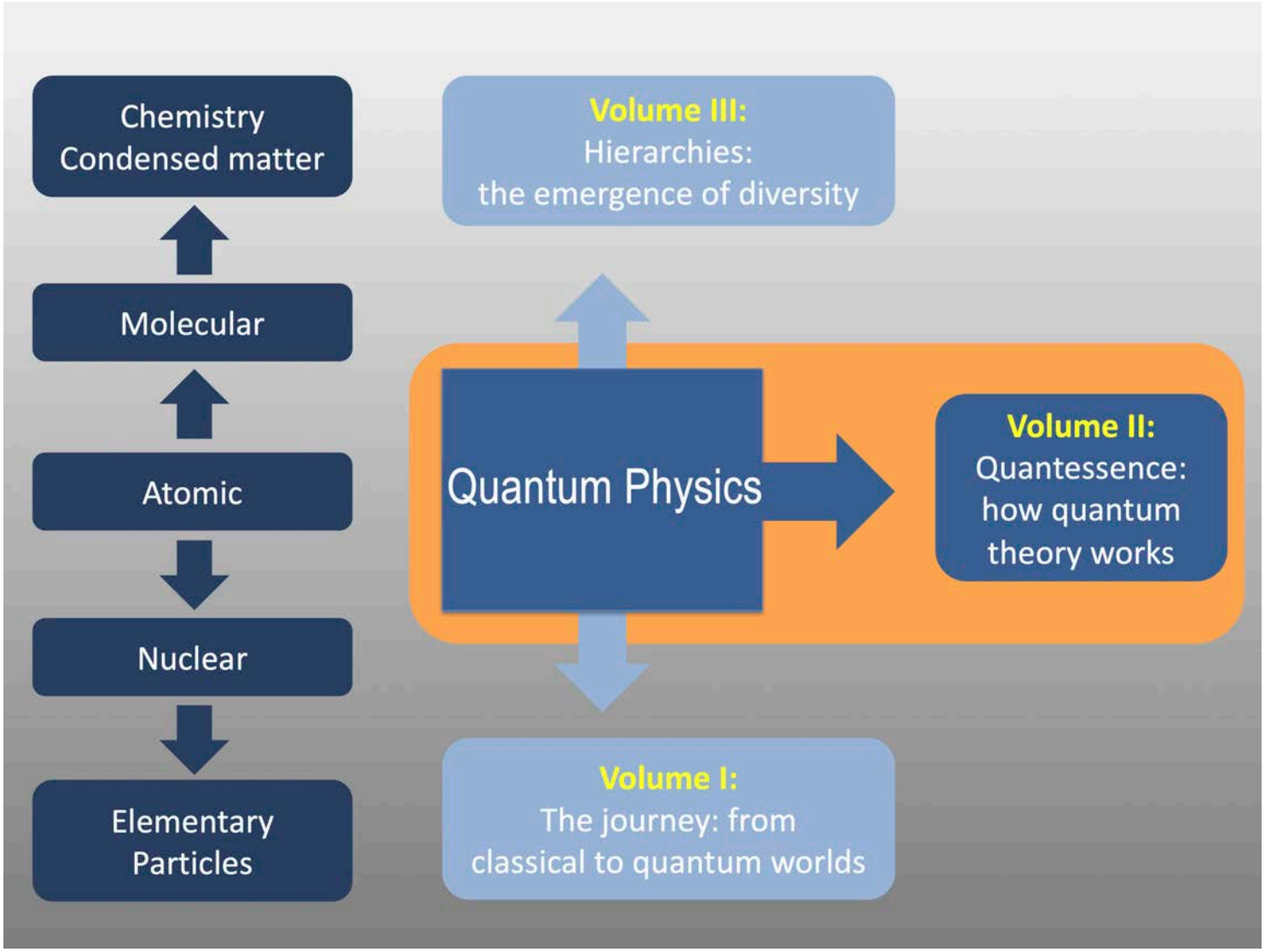
Amsterdam
University
Press

Quantessence: How Quantum Theory Works

In this volume we delve deeper into the mathematical structure underlying the theory. It focusses on the quintessence of quantum and is also the most conceptual. We introduce the space of states, the notion of observables, and cover subjects like qubits, entanglement, interference, and uncertainty relations. We reflect on the great paradoxes, the great equations, and their meaning.

Volume II

Quantessence:
how quantum theory works


Chemistry
Condensed matter
Molecular
Atomic
Nuclear
Elementary
Particles
Volume III:
Hierarchies:
the emergence of diversity
Quantum Physics
Volume II:
Quantessence:
how quantum
theory works
Volume I:
The journey: from
classical to quantum worlds

Contents

II Quantessence: how quantum theory works

Contents **239**

III Hierarchies: the emergence of diversity

The other side is usually a dark place?

Not necessarily. I think it has more to do with curiosity. If there is a door and you can open it and enter that other place, you do it. It's just curiosity. What's inside? What's over there? So that's what I do every day. [...,] once I start writing, I go somewhere else. I open the door, enter that place, and see what's happening there. I don't know–or I don't care–if it's a realistic world or an unrealistic one. I go deeper and deeper, as I concentrate on writing, into a kind of underground. While I'm there, I encounter strange things. But while I'm seeing them, to my eyes, they look natural. And if there is a darkness in there, that darkness comes to me, and maybe it has some message, you know? I'm trying to grasp the message. So I look around that world and I describe what I see, and then I come back. Coming back is important. If you cannot come back, it's scary. But I'm a professional, so I can come back.

The Japanese author Haruki Murikama in an interview by Deborah Treisman in The New Yorker (2019)

General references on quantum theory for Volume II:

- *Introduction to Quantum Mechanics*
  David J. Griffiths
  Pearson Education (2018)
- *Quantum Mechanics*
  Franz Mandl
  Wiley (2013)
- *Quantum Physics for Beginners*
  Carl J. Pratt
  Independent (2021)
- *The Feynman Lectures on Physics*
  R.P. Feynman (Author), R.B. Leighton (Contributor), M. Sands (Contributor)
  Pearson P T R; (3 Volume Set) 1st Edition (1970)
- *Quantum Mechanics: The Theoretical Minimum*
  Leonard Susskind
  Penguin Group (2017)
- *Principles of Quantum Mechanics*
  R. Shankar
  Springer (reprint of the original 1980 edition) (2013)
- *Foundations of Quantum Mechanics: An Exploration of the Physical Meaning of Quantum Theory*
  Travis Norsen
  Springer(2017)

Chapter II.1

The quantum formalism: states

> There's no sense in being precise when you don't even know what you're talking about.
>
> *John von Neumann*

Quantum theory has kept the community of physicists under its spell for over a century. It has opened new horizons for understanding a myriad of fundamental phenomena that were observed at ever deeper levels of nature, and it has produced a huge quantity of crucial results for the applied sciences. It has manifested itself in virtually all subfields of physics and from there entered into other adjacent fields like chemistry, engineering, informatics and even biology. And this process is still going on.
In this Volume we focus on the 'quantessential' features of the theory. This means that we will go into more detail with respect to the mathematical formalism underlying the theory. For pedagogical reasons we will apply it only to simple systems, and this may well give the impression that I am using a sledgehammer to crack peanuts.

The basic structure of the theory we are about to explore has far-reaching logical consequences. It will keep us busy in the following chapters on qubits, measurements, interference, entanglement and dynamics. We develop these concepts starting from the perspectives of classical physics, quantum physics and information physics. The starting point is always to define the system by the identification of its 'degrees of freedom' or basic dynamical variables. These can be 'external', like position, momentum, angular momentum or energy, or 'internal' where one may think of electric charge or something more exotic like intrinsic spin, isospin or color charge.

In Chapter II.1 we focus on the basic notions related to quantum states, such as state vectors, Hilbert space, separable versus entangled states, pure versus mixed states and the concepts of a density matrix and quantum entropy. In Chapter II.2 we discuss the notions of observables as operators, and the probabilistic nature of a quantum measurement. We also introduce the concept of incompatible observables, frames of reference and the Heisenberg uncertainty relations.
Chapter II.3 is about quantum interference in various double slit type of experiments, but also its manifestation in the so-called Berry phase.
In Chapter II.4 we turn to quantum teleportation and quantum computation. Teleportation is the consequence of the quantessential possibility of entangled states, which will be illustrated in a number of famous experiments and paradoxes. The results of recent experiments lead to the inescapable conclusion that quantum theory is correct. This means that theories built on hidden variables and local realism are no longer tenable in view of these experiments. Concerning quantum computation we introduce the notions of quantum gates and circuits, and discuss the factorization algorithm of Shor in some detail.

In Chapter II.5 we turn to the quantum theory of particles, fields and strings and illustrate a number of quantessential properties, such as the quantum statistics of particles and the spin-statistics connection. Volume II closes with Chapter II.6, where we give an overview of the role that symmetry and symmetry breaking play in physics and quantum physics in particular.

Quantum states: vectors in Hilbert space

If we describe a physical system in the classical realm, the relevant variables like position, velocity or momentum and energy are part of the definition of the system. They are observables in that we can measure them, thereby producing dimensionful values as an outcome.

We have mentioned what in quantum physics the states look like: they are vectors in some rather abstract state space called the Hilbert space, and in this section we will show how and to what extent the ordinary physical variables can be retrieved from the state vector.

The crucial fact is that in the quantum formalism observables are not represented by just numbers but are defined as *matrices* or *operators* acting on the state space. That sounds complicated, and yes, it is. It illustrates a remark made by Paul Dirac who stipulated that matters, which at a certain moment may be considered merely as pastimes for mathematicians and logical thinkers, may turn later into tools that are indispensable for understanding nature. And if understanding nature is our goal it may be worthwhile to familiarize ourselves with these mathematical concepts, just like the pioneers of quantum theory had to do a century ago.

In this chapter we point out the quantessential differences between classical and quantum systems for the simplest of all quantum systems, the *quantum spin* or *qubit*. This two-level system plays a fundamental role in many applications of quantum theory, but is also a favorite toy-model. The ability to control and manipulate arrays of qubits is the holy grail of quantum technology as it entails the production of quantum information processing devices that enable for novel applications, varying from quantum key distribution and teleportation to quantum computation. It is a major challenge to find physical implementations of a basic qubit that can be reliably manipulated and at the same time can be scaled to large arrays.

Reader alert. Remarkably, in talking about quantum concepts and meaning, formulas are often easier to understand than words. However, if you are not familiar with the notion of operators and matrices, don't despair! The philosophy of the book is not to shy away from them, but to plug and play with them in the simplest imaginable cases to gain familiarity with them. As with driving lessons, you don't have to drive all the way from Spokane to Miami Beach and back to get a proper appreciation for what a highway is. I kindly request that you accept the definitions for what they are, then we will play around a bit so that you will end up throwing matrices around like ordinary numbers.

I will supplement the rather abstract algebraic language of matrices and the like, whenever possible, by more geometric images; for most people imagery provides more insight and is easier to remember. And talking about vectors and matrices, I should like to remind you of the respective *Math Excursions* at the end of Volume III, because those intros will make understanding the forthcoming chapters a lot easier. The use of a symbolic language will at least keep us from slowly getting lost in a dense fog of ever more cryptic quantum terminology and quantum vagueness. Take my word, or rather, my equations for it. □

This challenge is approached from many different angles, like quantum optical systems, superconducting devices, atoms in optical latices, ions in traps, and topologically ordered phases. Progress is rapid which means that quantum devices exploiting the fundamental features of quantum theory may well be with us in a decade or two.

Quantum versus classical

> I think it is safe to say that no one understands quantum mechanics. Do not keep saying to yourself if you can possibly avoid it: 'But how can it be like that?' because you will go down the drain 'into a blind alley from which nobody has yet escaped. Nobody knows how it can be like that.'
>
> *Richard Feynman*

We start by comparing the quantum and classical world generally. The fundamentally different concepts and formulations have profound consequences for the logical and deductive structure of the theories. Where do these worlds meet or separate? Actually, do they?

Classical systems. In classical physics it is usually quite obvious what the system consists of and what the possible states are. If we talk about a *particle* for example we will typically specify the state by assigning it a mass m, a position x, and a velocity v. Given the state at some initial time, Newton will tell us what the state will be at any later time, provided we know the forces that act on the particle along the way. For a *field* like the electromagnetic field we specify the field configuration, by which we mean that we give the electric E and magnetic B fields over all of space. Then the Maxwell equations tell you all about the time development of that initial field configuration, provided we know what the external charges and currents, usually called *sources*, are. The evolution of the gravitational field is described in a similar way by the Einstein equations. Subsequently we have to combine the frameworks of Newton, Maxwell and Einstein to get the actual time development of the complete classical system of particles with and without charge and gravitational and electromagnetic fields. The structure of the theory is absolutely unambiguous, based on a clear methodology.

Yet, the coupling of the different components of fields and sources makes the system extremely nonlinear and therefore hard to solve explicitly. For example there is the intricate problem of the 'back reaction': the fields will not only change as a consequence of the movement of the charges, but in addition the accelerated charges will radiate. There are certain simple cases that can be dealt with analytically through closed expressions in terms of standard functions, but mostly that is not the case. Whereas we can solve the Newtonian two-body problem analytically, this is not the case for the three-body problem. One has to resort to numerical procedures which can become extremely cumbersome, if one insists on high accuracy, which is the case if one wants to make predictions about the behavior of the system on long time-scales. This point leads us to an additional observation that should be made concerning classical physics.

Nonlinear dynamics and deterministic chaos. We just stated that if we know for example the position and velocity of a particle at a given instant in time, the time evolution is completely fixed by Newton's laws provided we know the forces acting on the particle. This implies that any uncertainty in its evolution is driven by the limited accuracy of the initial conditions. This is not as innocuous as it sounds even if one has a huge zoo of advanced computers at one's disposal. What we have learned in the last half century from studying simple nonlinear systems is that already on a classical level, such systems – *in spite of being completely deterministic* – can exhibit chaotic behavior. In such situations it is not possible to make precise long-term predictions, because small initial uncertainties can be amplified exponentially in time by the chaotic dy-

namics of the nonlinear system. These systems exhibit an extreme sensitivity on initial conditions often referred to as the *butterfly effect*, meaning that a tiny change in the initial condition may lead to vastly different consequences a relatively short time afterwards. However, what concerns us here is that within classical physics there is no fundamental limit on the accuracy of measurements – by measuring more and more carefully, we can predict the time evolution of a system more and more accurately. The system is fundamentally deterministic. This is no longer true in the quantum world because there we will run into a fundamental limit on the accuracy of the simultaneous measurement of physical observables.

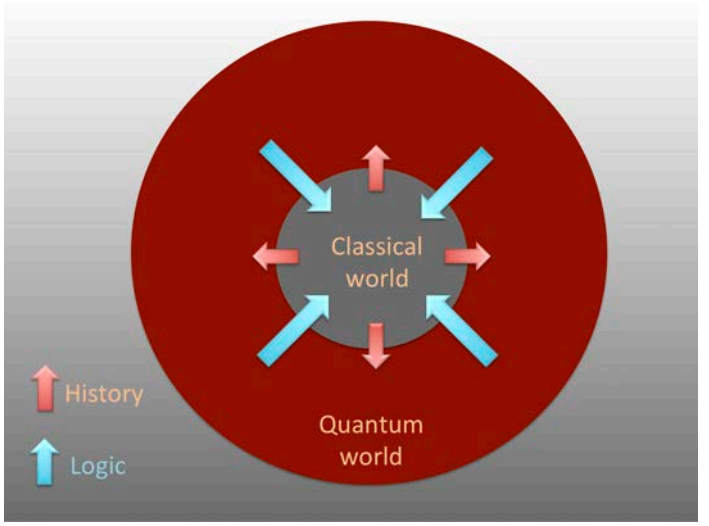


Figure II.1.2: *Classical versus quantum.* We were born in a classical world, but after exploring the nature of things we have discovered the existence of a much larger quantum world. Once these discoveries were made, we understood that the logic should be reversed: it is the classical world that can be logically deduced from the quantum world, and not the other way around.

The correspondence principle

Where classical and quantum meet. At the most basic level there are fundamental differences between the classical and the quantum frameworks. On macroscopic scales, meaning relatively large scales of space, time and energy, where we know classical physics works well, the predictions of classical and quantum theories of course have to agree. This requirement is known as the *correspondence principle*. There is no logical path that brings you from classical physics to quantum physics, but the converse is certainly possible and even mandatory. We should insist on understanding the emergence of all of classical physics from the underlying quantum description. This turns out not to be straightforward at all, but then, nobody promised us it would be. In Figure II.1.2 we have symbolically indicated the classical and quantum worlds. We contrast the direction of the historical process of scientific evolution, moving us out of the classical into the quantum domain, versus the direction of logical deductions and implications which go the opposite way. It warns us that we should not strive for an interpretation or representation of quantum content in classical terms, that would be a terribly misguided effort indeed. So, historically, quantum theory emerged out of the classical theories, but logically it is the other way around, and that is inherent to the way knowledge transcends itself in the process of scientific progress.

Classical phenomena with quantum explanations. As we discussed in the previous Volume, for example in Chapter I.2 , the scale of the quantum regime is set by Planck's constant h , or *h-bar* defined as $\hbar \equiv \mathrm{h}/2\pi$, which has dimensions of $\mathrm{energy}\times\mathrm{time}$ (or equivalently $\mathrm{momentum}\times\mathrm{length}$). Because of the tiny value of this constant, we expect the quantum properties to become manifest at small time and length scales, and low temperatures. However, collective macroscopic behavior is to a large extent an indirect manifestation of the properties of the basic constituents of the system, and of the interactions between them and the environment. After all, not withstanding the striking similarities between an ant colony and human society, the

even more striking differences between them can be largely traced back to the differences between an individual human being and an individual ant. Looking at matter in a similar way, one expects that radically different properties at a microscopic scale (say at the level of atomic and molecular structure) may in turn lead to fundamentally different collective behavior of these basic building blocks and therefore to different emergent properties on a macroscopic scale. So, one certainly should expect quantum manifestations on a macroscopic scale after all. Indeed, most phases of condensed matter realized in nature, such as crystals, ordinary conductors, semiconductors, superconductors or magnetic materials, all involve forms of collective behavior that can only be understood from a quantum perspective. The (meta-)stability and structure of matter is intimately linked to the quantum behavior of its fundamental constituents.

The quantum domain. Returning to the question of states, as we will see in this and the following chapters, the quantum states of bits, particles or fields are very different from their classical precursors and in the beginning it was even far from evident what the space of states would be. However, once we found out, we learned that the structure of the state-space tells us a lot about the generic features of quantum systems and how these may radically differ from their classical analogues. Studying the underlying mathematical structure will enable us to anticipate what we might expect in real physical situations. With some exaggeration one could say that everything that is not forbidden is compulsory, and henceforth will manifest itself somewhere in Nature. Nature *is* quantum.

Many exotic quantum features like particle interference or entanglement derive directly from its underlying structure, but that didn't make it any easier to demonstrate these features through experiment. Many predictions of quantum theory have lingered on the margins, waiting for experimental techniques to develop to the required level of precision. There are quite a few examples where it has taken more than half a century before predictions could be put to the test. Science requires not only brilliance but also patience. Nowadays, many quantessential phenomena can be beautifully demonstrated by experiments exploiting superconductivity and quantum optics. There is still much more to discover, which is why we want to explore these quantum state spaces and their remarkable properties in this separate second volume. Whereas the present state of modeling real systems in nature within the quantum mechanical framework is described in the Volumes I and III, this volume is dedicated to the 'cosmic code' itself.

Classical states: phase space

The state at some time t *of a classical system is specified by assigning values to a minimal subset of dynamical variables from which all possible other variables can be calculated. We say that the state of the system corresponds to a point in phase space* $\mathcal{F}_{\mathrm{ph}}$ *. We are going to discuss the case of a basic particle and work out the discrete 'Newtonian' dynamics of an Ising spin or classical bit as an example.*

Phase space. To specify the state of a simple particle, which may have a mass m and a charge q , we have to give its position $\mathbf{x}$ and its velocity $\mathbf{v}$ or momentum $\mathbf{p} = \mathrm{m}\mathbf{v}$. The space of positions is usually called *configuration space* and denoted as $\mathcal{X}$. In three-dimensional space both position and velocity have three components because they are $vectors$, and thus the *phase space* $\mathcal{F}_{\mathrm{ph}} \simeq \{\mathbf{x}, \mathbf{p}\}$ has six dimensions. From the point of view of particle dynamics, mass and charge are just fixed external parameters. Note that other dynamical variables of a particle, like its energy or angular momentum, can be expressed in terms of velocity and position and therefore can be calculated once the point in phase space is given.

A *property* corresponds to a subspace of the phase space.

A state of the system can be assigned a property, in the sense that one can decide whether a property is true or false by determining whether the point representing the state of the system is lying in or outside that subspace.

The dynamical system will develop in time according to some dynamical equations like Newton's equations of motion, and the point describing the state will move in phase space correspondingly. Furthermore in classical physics it is assumed that the point can in principle be determined to arbitrary precision by a simultaneous measurement of the basic variables thereby fixing the point in phase space. And one also assumes that observations can be made which do not disturb the system, and hence do not affect the trajectory in phase space. These assumptions are an essential Volume of the classical physics paradigm.

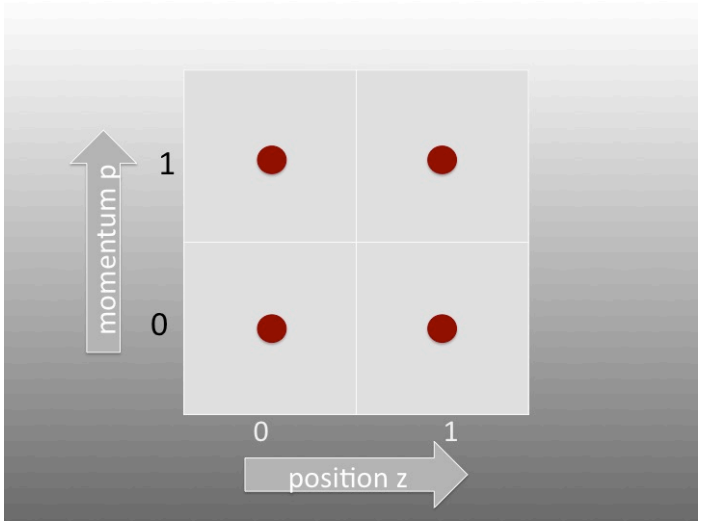


Figure II.1.3: *Phase space.* The phase space of the dynamical bit consists of four points.

The mechanics of a bit

Let us now turn to a system even simpler than a single particle, which I call a *dynamical bit*. We are going to do a bit of bit mechanics. I have chosen this system because it links basic classical mechanics to basic information theory, and defines a simple quantum system as well. As we all know, a bit has two states (positions) labeled $z = 0$ and $z = 1$, so its configuration space consists of two isolated points. Introducing a discrete time step (like the clock in a computer) allows us to define a *discrete dynamics*. We distinguish two possibilities: after the time step the bit changed to the other state or it stayed where it was. This begs for an additional binary state variable which we appropriately call the *bit-momentum* p. So its value labels two distinct states of motion, where $\mathrm{p} = 0$ means 'at rest' or $\mathrm{p} = 1$ meaning 'on the move.'

Both the classical position and the classical momentum space consist of two points, and therefore both bit-position and bit-momentum are binary variables, which means that all values can be added $\mathrm{mod}\ 2$, meaning in particular that $1 + 1 = 0$.

Binary mechanics. The phase space for this dynamical bit corresponds to four points

$$\mathcal{F}_{\mathrm{ph}} \simeq \{\mathrm{p}, \mathrm{z}\} = \{0, 0; 0, 1; 1, 0; 1, 1\}$$

as indicated in Figure II.1.3. To push the comparison with Newtonian mechanics even further, one could say that the dynamical state in the absence of further interactions would be characterized by the conservation of momentum. Then with $\mathrm{p} = 0$ the bit would be 'at rest' indefinitely, in which case the position is conserved as well, but with $\mathrm{p} = 1$, the bit stays constantly hopping between the two position states. Depending on the initial condition one finds two fixed points and one two-cycle. The phase space picture of the possible dynamics is given in Figure II.1.4 (top). Maybe you have already noted the amusing possibility of introducing a *bit-force* F, defined *à la* Newton as the change in bit-momentum. Also F takes a binary value; $\mathrm{F} = 0$ leaves the momentum unchanged, while with $\mathrm{F} = 1$ the momentum value changes, which leads to a different dy-

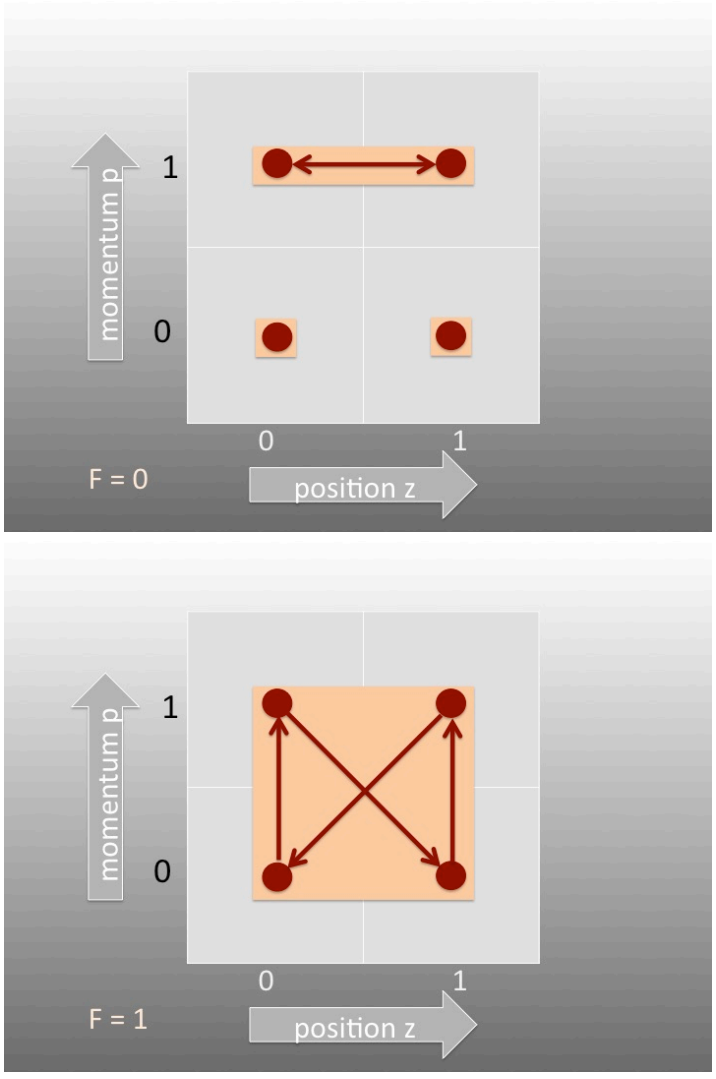


Figure II.1.4: *Bit mechanics.* Phase space picture of 'Newtonian' bit-dynamics with a binary force F being either 0 (top) or 1 (bottom). For $\mathrm{F} = 0$ there are two fixed points and one two-cycle, for $\mathrm{F} = 1$ there is only one four-cycle.

namic consisting of the four-cycle depicted in Figure II.1.4 (bottom). These variables are elements of a *Boolean algebra*, discussed in the *Math excursion* on algebras in Volume III.

Complementary representations. The system is clearly completely deterministic, because given the initial binary z and p values, its future states after an arbitrary number of time steps can be calculated. These discrete dynamics are like a little automaton, an updating procedure for the z-bit that depends on the p-bit. Updating means that the states of the two-bit system change and therefore the dynamics define a logical gate in the sense of digital computation. So we have arrived at four alternative ways to characterize the dynamics of the bit:
(i) as an *updating algorithm* or *iterative map* $|\mathrm{in}\rangle \to |\mathrm{out}\rangle$,
(ii) as a *diagram* representing the gate,
(iii) as a two-bit to two-bit *input-output table*,
(iv) and as a 4×4 matrix acting on the column vector of two-bit in-states $(\mathrm{p}, \mathrm{z}) = \{0, 0; 0, 1; 1, 0; 1, 1\}$.
For $\mathrm{F} = 0$ this looks as follows: (i) the algorithm generating the dynamics is just,

$$(\mathrm{p}, \mathrm{z}) \to (\mathrm{p}, (\mathrm{z} + \mathrm{p}) \bmod 2)\,,$$

which corresponds to the (ii) diagram, (iii) state map, or (iv) the (block-diagonal)matrix as given in Figure II.1.5.

Gates and information dynamics. From the picture we learn that the two-bit dynamic is in fact generated by a two-bit gate which is well known as the *controlled* NOT- or CNOT-gate. The diagram should be read as follows: the horizontal lines correspond to the two incoming (left) and outgoing (right) bits. It is a conditioned gate, which is indicated by the vertical line from the p-bit to the z-bit. The encircled plus symbolizes a NOT-gate acting on the z-bit, but its action is conditioned on the value of the p-bit: it is activated if $\mathrm{p} = 1$ and not if $\mathrm{p} = 0$. The dot on the p-line indicates that it is the control bit, not changing value by passing the dot. With this interpretation it is straightforward to compute the entries of the input-output table. One puts the input state on the lines at the left and then follows the lines through the diagram to the right performing the instructions one encounters.

This matrix acts like a permutation matrix on the input column vector of two-bit in-states. Indeed, we see that on the

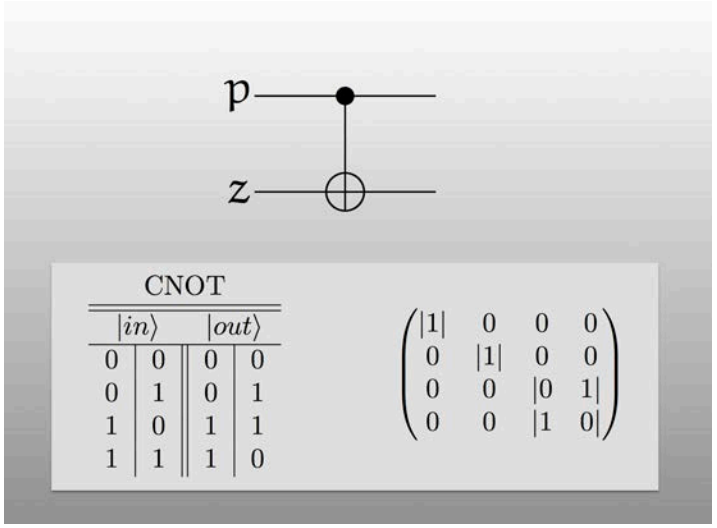


Figure II.1.5: *Three representations.* The $\mathrm{F} = 0$ bit dynamics is generated by the CNOT-gate. In the 'block-diagonal' matrix representation on the right, we marked the two fixed points and the two-cycle.

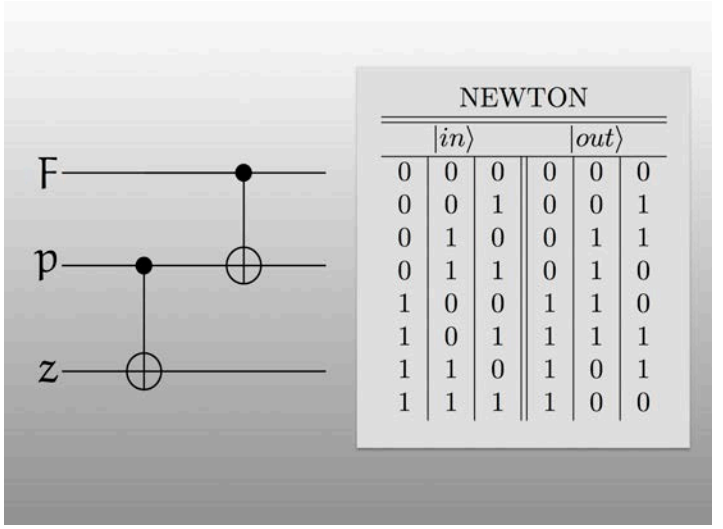


Figure II.1.6: *NEWTON-map.* The three-bit *NEWTON-gate* and the corresponding $|\mathrm{in}\rangle \to |\mathrm{out}\rangle$ map acting on the column vector of $(\mathrm{F}, \mathrm{p}, z)$ states.

top two entries it acts like a unit matrix, while on the bottom two entries (x) it acts like a NOT-gate.

The NEWTON gate. Imagine that we also include the 'bit-force' we defined as a third force-bit F. Then we obtain an interesting three-bit gate for the complete dynamics of the system. One finds that it can be characterized by the updating algorithm:

$$(\mathrm{F}, \mathrm{p}, z) \to (\mathrm{F}, (\mathrm{p} + \mathrm{F}) \bmod 2, (z + \mathrm{p}) \bmod 2)\,,$$

which corresponds to the diagram and state map of Figure II.1.6 and the matrix in equation (II.1.1).

On the first four rows it acts like a CNOT, and in the second block it performs some sequence of permutations. In that sense this NEWTON-gate actually computes something on three bits, but from the diagram we see that it is not an irreducible three-bit gate, rather it is composed of two sequentially applied CNOT-gates. It has the following 8×8 matrix structure in a basis given by the first three columns of the $|\mathrm{in}\rangle$ states of the table in Figure II.1.4. Note that due to the four bottom entries corresponding to $\mathrm{F} = 1$, the fourth power of the NEWTON-gate is equal to the unit matrix. Hence, the dynamics generated has indeed period four, as one would expect if the force is constant. That causes p to hop with period two and z with period four. It is the dynamics of the bottom diagram in Figure II.1.4.

$$\text{NEWTON}: \begin{pmatrix} |1| & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & |1| & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & |0 & 1| & 0 & 0 & 0 & 0 \\ 0 & 0 & |1 & 0| & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & |0 & 0 & 1 & 0| \\ 0 & 0 & 0 & 0 & |0 & 0 & 0 & 1| \\ 0 & 0 & 0 & 0 & |0 & 1 & 0 & 0| \\ 0 & 0 & 0 & 0 & |1 & 0 & 0 & 0| \end{pmatrix} . \quad \text{(II.1.1)}$$

The matrix corresponding to this NEWTON-gate, displayed above, is unitary in the sense that the transpose of the matrix is indeed its inverse. But the matrix is not symmetric, meaning that it is not a time reversal invariant process, be-

cause then the matrix would have to be its own inverse. This, however, *is* the case for the CNOT-gate represented by the matrix in Figure II.1.5 .

Conserved energies. In classical Hamiltonian mechanics one may derive the equations of motion, or the time evolution once the energy function is given, as we showed in Chapter I.1 . In the case of discrete dynamics it is less straightforward as we cannot take derivatives in the normal way. Because all variables are binary, small variations are nonexistent! The role of the Hamiltonian is played by the updating algorithm because that generates the time translation of the system. It is that mapping, which by repeated application maps out the time trajectory of the system in phase space. In these discrete cases one may invert the question by asking whether there is a (binary) energy function $\mathrm{E}(\mathrm{p},z)$ that is conserved in the time series, i.e. whose value does not change for the subsequent points on a given orbit in phase space.

Let us look at some simple candidates. These can come across as slightly unusual, exactly because the energy is also a binary variable, implying that it can take only two possible values. The good thing about that is that the energy stays always bounded and therefore the system is always well-defined.

Example 1: $\mathrm{E}=\mathrm{p}$. You would expect the energy of a free particle to be proportional to p^2, and since p is Boolean variable we have that $\mathrm{p}^2=\mathrm{p}$. The free particle does not experience any force and so one expects that the Newtonian dynamics rule $(\mathrm{p},z)\to(\mathrm{p},z+\mathrm{p})$ will apply. This is indeed the case where $\mathrm{F}=0$ which we discussed before and p is preserved. It has two fixed points with $\mathrm{E}=0$ and one periodic orbit of length two with $\mathrm{E}=1$:

$$(0,0)\ ;\ (0,1)\ \ \mathrm{and}\ \ (1,0)\leftrightarrow(1,1)$$

Example 2: $\mathrm{E}=\mathrm{F}=1$ This is the case of a non-zero constant force conserved under the Newtonian rule $(\mathrm{p},z)\to(\mathrm{p}+1,z+\mathrm{p})$. Its action corresponds to one periodic orbit of length four with energy $\mathrm{E}=1$.

$$(0,0)\to(1,0)\to(0,1)\leftrightarrow(1,1)\to(0,0)\to\cdots$$

Example 3: $\mathrm{Q}=\mathrm{p}+z$. The function Q is a conserved 'charge', or 'constant of the motion' under the clearly not Newtonian rule $(\mathrm{p},z)\to(\mathrm{p}+1,z+1)$. Again it has two periodic orbits of length two which are now along the diagonals of the phase space, one with $\mathrm{E}=0$ and the other with $\mathrm{E}=1$:

$$(0,0)\leftrightarrow(1,1)\ \ \mathrm{and}\ \ (1,0)\leftrightarrow(0,1)$$

Quantum states: Hilbert space

We discuss the generic setting of a quantum system. For a quantum system we have a set of states denoted $\{|\Psi\rangle\}$, *which are vectors that correspond to elements of the so-called Hilbert space* $\mathcal{H}$ *of the system. The basic quantum setting introduces two novel ingredients, one is the* complexification*, and the other the* linear superposition principle *of states. These have dramatic consequences.*

The Hilbert space of states. To explain the basic ideas of quantum theory, or for that matter of quantum information, we will in this section restrict our attention again mainly to the *qubit*, which can be viewed as the basic building block of quantum information systems. The physical state of a quantum system is described by a wavefunction which can be thought of as a vector in an abstract multidimensional space of states, called the *Hilbert space* denoted by $\mathcal{H}$. For the moment, this is just a finite dimensional vector space where the vectors have complex, rather than real, coefficients, and where the length of a vector is the usual length in such a space, i.e. the square root of the sum of the (absolute) squares of its components along the axes.

Hilbert space replaces the concept of phase space in classical mechanics. Collections of observables, or measurable variables such as spin, charge, position, or momentum, can be used to set up an orthogonal basis for the Hilbert space.

As we will see, a dramatic difference from classical mechanics with tremendous consequences is that many quantum mechanical quantities, such as position and momentum, or spin components along the x-axis and the y-axis, *cannot* be measured simultaneously. Another essential difference from classical physics is that the dimensionality of the state space of the quantum system is huge compared to that of the classical phase space. To illustrate this drastic difference, think of a particle that can move along an infinite line with an arbitrary momentum. From the classical perspective it has a phase space that is two-dimensional and real (a position x and a momentum p), but from the quantum point of view the particle is described by a wavefunction Ψ of one variable (typically the position x or the momentum p). The state is thus determined by specifying a function for all points x. As the state corresponds to a function, the state space must be a 'space of functions.' Formally such a wavefunction corresponds to an element of an infinite-dimensional Hilbert space which is a space of functions that satisfy certain restrictions. So, we go from two real numbers classically to a complex function of one variable in the quantum domain. That is quite a difference indeed! We will address the topic of quantum particles in detail in Chapter II.5.

States of a quantum bit

Now you might have thought that this is not such a big deal, because the classical state corresponds to a point in phase space and that point can be characterized by a vector in phase space. But this is not the way to think about it. We just mentioned the dynamical bit as an example of an almost trivial dynamical system. To this classical system corresponds a quantum system called the quantum bit or *qubit* for short, and the statement is that to every point in the configuration space of the classical bit we associate a basis vector of the Hilbert space. So the bit-position space consists of two points $\{1, 0\}$, and hence the Hilbert space of the qubit is two-dimensional and may be thought of as spanned by two orthogonal unit vectors $\{|1\rangle, |-1\rangle\}$.[1]

A general state of a qubit is described by a wavefunction or *state vector* $|\psi\rangle$, also called a *ket* or *ket vector*, which can be written as

$$|\psi\rangle = \alpha|+1\rangle + \beta|-1\rangle \text{ with } |\alpha|^2 + |\beta|^2 = 1\,, \qquad \text{(II.1.2)}$$

where α and β are complex numbers.[2] *Any* linear combination of the two basis states corresponds to an admissible quantum state, as long as it satisfies the *normalization condition*, meaning that the sum of the squares of the components equals one. This means that you can think of $|\psi\rangle$ as a unit vector in the 2-dimensional complex vector space, denoted $\mathbf{C}^2$ spanned by the two basis vectors $|1\rangle$ and $|-1\rangle$.

The geometry of qubit state space. What we have learned so far is that a finite state classical system will lead to a finite-dimensional complex vector space for the corresponding quantum system. Let us describe the geometry of the quantum configuration space of a single qubit in more detail. The constraint $|\alpha|^2 + |\beta|^2 = 1$ says that the state vector has unit length, which defines the complex unit circle in $\mathbb{C}^2$, but if we write the complex numbers in terms of their real and imaginary parts as $\alpha = a_1 + ia_2$ and

[1] We switch from a '$1, 0$' labeling in the classical domain, to a '$1, -1$' labeling in the quantum domain, these are matters of notation and of mathematical convenience as we will see later.

[2] For a tailor-made introduction to complex numbers and vectors see the *Math excursions* on pages 630 and 632 of Volume III. It is important for complex numbers that basic algebraic operations like addition, subtraction, multiplication and division can be defined. It is almost like in the musical *Annie get your Gun*: 'Everything you can do I can do better.'

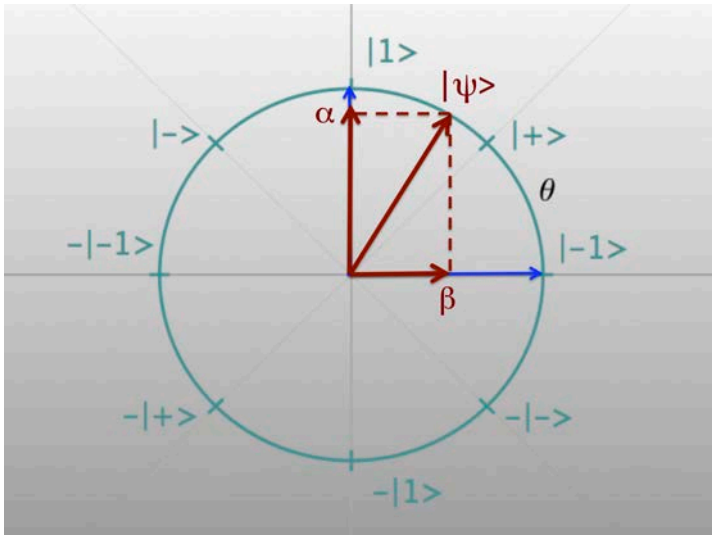


Figure II.1.7: *State decomposition.* Decomposition of a real qubit state vector $|\psi\rangle$, the purple arrow, into its components α and β with respect to the blue basis or frame $\{|+1\rangle, |-1\rangle\}$. The circle represents the subspace of the real states, and in that case we clearly have that $\alpha = \sin\theta$ and $\beta = \cos\theta$. We have marked some of the other real states that we will refer to in the text.

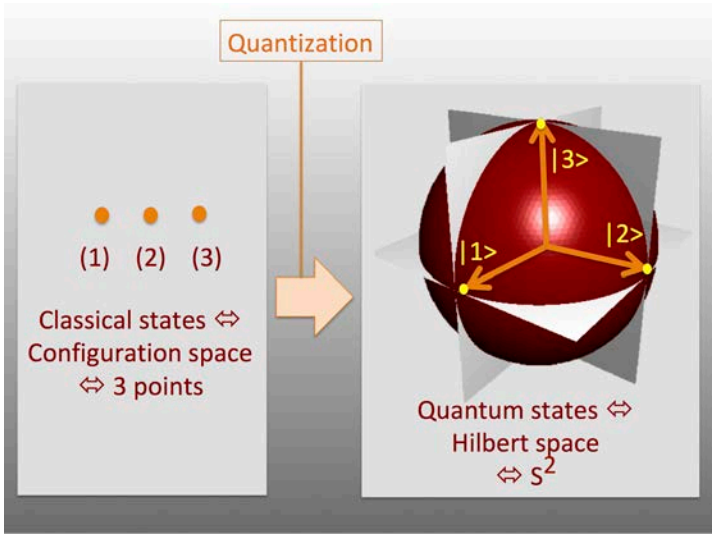


Figure II.1.8: *Configuration versus Hilbert space.* A classical system with a configuration space corresponding to a set of three points. The quantum Hilbert space for this system would correspond to the unit-sphere in the complex three-dimensional space $\mathbb{C}^3$. In the figure we show the restriction of that space to real states forming a two-sphere. Classical and quantum spaces are structurally very different. There is a 'world' in between which is described by the formalism we are about to explore.

$\beta = b_1 + ib_2$, then we obtain $|a_1 + ia_2|^2 + |b_1 + ib_2|^2 = a_1^2 + a_2^2 + b_1^2 + b_2^2 = 1$. The geometry of the space described by the latter equation is just the three-dimensional unit sphere S^3 embedded in a four-dimensional Euclidean space, $\mathbb{R}^4$ with coordinates a_1, a_2, b_1, and b_2. This three-dimensional sphere is in physics referred to as the *Bloch sphere*.

Complex rotations. At this point it is appropriate to make a side comment. As the state of a qubit is a normalized two-dimensional complex vector, the state space of a qubit corresponds to a complex circle, which in turn equals S^3. All states on the complex unit circle can be obtained by acting with all complex rotations on a given qubit state in $\mathbf{C}^2$. This is by definition the group $SU(2)$ and having argued that these vectors can be transformed into each other by the elements $U \in SU(2)$, we can also conclude that the space of all $SU(2)$ transformations is in one-to-one correspondence with the points on the three-sphere S^3. We will use these geometric representations of state spaces and transformation groups later on, because they are easier to understand than just formulas.

Real states. For pedagogical reasons it is advantageous to limit ourselves for the moment to the subspace corresponding to real states. This means that one only considers states for which α and β are real and the condition $\alpha^2 + \beta^2 = 1$ imposes that the states lie on an ordinary circle in $\mathbb{R}^2$. The real states are depicted in Figure II.1.7, where we have also marked some special states. Many of the formal quantum properties can be explained within this real subspace.

Alternative notations. We may represent the state by the column vector of its components:

$$|\psi\rangle \Leftrightarrow \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

If you like you can also map the states of the classical configuration space in the quantum picture, then the classical bit would only have the two states $|\pm 1\rangle$, corresponding to the basis vectors

$$|\pm 1\rangle \Leftrightarrow \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

while the qubit can be any normalized linear combination of these two basis states. This makes the dramatic difference between the classical and quantum setting quite visible indeed. Each point in the configuration space $\mathcal{Z}$ of the classical system corresponds to an orthogonal basis vector of the Hilbert space, and consequently adding a point to the configuration space $\mathcal{Z}$ adds a dimension to $\mathcal{H}$. So in this picture the classical states correspond to the corners of a unit hypercube in that higher dimensional space, while the quantum states lie on the unit-hypersphere embedded in that space. This is illustrated in Figure II.1.8 for a three-state system.

The scalar or dot product

Ordinary, say 'high school' vectors are called *real* vectors. You may remember how the length $|a|$ of a vector $\mathbf{a}$ was defined as the square root of the sum of the squares of its components $|a| = \sqrt{a_1^2 + a_2^2 + \ldots}$. And the dot or inner product of two vectors $\mathbf{a}$ and $\mathbf{b}$, wassimilarly as $\mathbf{a} \cdot \mathbf{b} = \sum a_1 b_1 + a_2 b_2 + \ldots = |a||b| \cos\theta$, with θ the angle between them.

Conjugate states. For the state or ket vectors $|\psi\rangle$, we basically want to do the same thing, but because the vectors are complex, it is slightly more complicated. However, once you understand the definition, a notation introduced by Dirac will make it like 'real' vectors. We first define the *dual* of the vector space in $\mathbf{C}^2$ with dual or conjugate vectors, called *bra vectors*, that can either be represented as row vectors with complex conjugated elements, where $\alpha^* \equiv a_1 - i a_2$ etc. Following the notation introduced by Dirac we write this like,

$$\langle\psi| = \langle 1|\alpha^* + \langle -1|\beta^*. \qquad \text{(II.1.3)}$$

This somewhat strange nomenclature of *bra* and *ket* vectors makes more sense once you realize that they allow you to make a *bracket*, and this bracket is nothing but a scalar product of two vectors.

The inner product The *scalar (or inner, or dot) product* maps a bra-and-ket-pair into a complex number (the scalar). So if we have two state vectors $|\psi\rangle$ and $|\phi\rangle = \gamma|1\rangle + \delta|-1\rangle$ then their bracket is defined as

$$\langle\phi|\psi\rangle = \langle\psi|\phi\rangle^* = \gamma^*\alpha + \delta^*\beta. \qquad \text{(II.1.4)}$$

As the components of the state vectors are complex, the dot product of two vectors is also, and it is thus no longer true that it equals the product of the lengths of the vectors and the cosine of the angle between them. But, just like in the real case, we call two vectors whose dot product vanishes *orthogonal* or *perpendicular*. Similarly, the inner product of a vector with itself, which is always a real number, is defined as the length squared of that vector.

Probability amplitudes. It turns out that the dot product of state vectors has an important physical interpretation as a *probability amplitude*, and it plays a fundamental role if we are going to talk about quantum measurements. We will discuss this extensively later in this chapter, but it is useful to preview here already the basic idea. Let us look at Figure II.1.7, where we have a state $|\psi\rangle$, and if we want the outcome with respect to the blue $\{|1\rangle, |-1\rangle\}$ frame, then the probability to find the outcome $+1$ would be the

probability amplitude squared:

$$p_{+1} = |\langle 1|\psi\rangle|^2 = \langle\psi|1\rangle\langle 1|\psi\rangle = \alpha^*\,\alpha = |\alpha|^2\,. \qquad \text{(II.1.5)}$$

This assignment of a probability to the inner product of two state vectors is called the *Born rule*, after Max Born, the quantum pioneer who proposed the probability interpretation of quantum mechanics. It is also referred to as the *Kopenhagener Deutung*, or *Copenhagen interpretation*. Clearly a similar calculation for the -1 outcome would give the probability $p_{-1} = |\beta|^2$. The normalization of the state vector is just the statement that the total probability for finding one of the two possible outcomes is one: $p_{+1} + p_{-1} = 1$. Making a measurement means that we get new information on the state and that affects the probabilities for the measurements after that. This means that the state vector has to change, because it has to reflect the probabilities of measurement outcomes at any instant. In this simple example the following happens, if we obtain $+1$ the state will change to the plus one state: $|\psi\rangle \rightarrow |1\rangle$. So the state gets 'projected' on the state, which gives that measurement outcome with unit probability. This you can interpret as saying that if you measure a quantum system and find a certain outcome, then if you repeat the measurement immediately afterwards you will find the same outcome.

Projectors. There is an alternative way to read equation (II.1.5). One needs to first look at the object,

$$P_1 = |1\rangle\langle 1|\,; \qquad \text{(II.1.6)}$$

this is not an inner product, but rather a *projector* on the state $|1\rangle$. If this operator acts on an arbitrary state $|\psi\rangle$, it produces the projection equal to $\langle 1|\psi\rangle$, along the $|1\rangle$ basis vector:

$$P_1\,|\psi\rangle \rightarrow |1\rangle\langle 1|\psi\rangle\,.$$

So the probability to find an outcome $+1$ is also obtained by 'sandwiching' the projector P_1 in the state $|\psi\rangle$:

$$p_{+1} = \langle\psi|\,P_1\,|\psi\rangle\,.$$

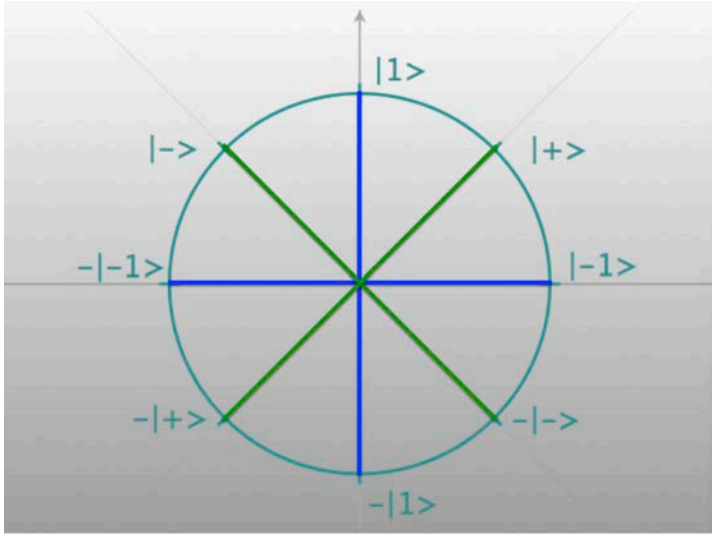


Figure II.1.9: *Two frames.* We have depicted two different frames for the two-dimensional qubit Hilbert space. The blue basis consists of the states $\{|\pm 1\rangle\}$, whereas the green basis consists of the states $\{|\pm\rangle\}$.

These probability and measurement definitions will be used extensively in the next chapter.

A frame or basis

It is convenient to choose an *orthonormal frame* consisting of unit length, mutually orthogonal basis vectors that 'span' the vector space. This amounts to choosing a set of basis vectors $|i\rangle$ with $i = -1, 1$, which have the property that:

$$\langle i|j\rangle = \delta_{ij}\,, \qquad \text{(II.1.7)}$$

with the Kronecker 'delta' symbol defined as follows: δ_{ij} equals one if $i = j$, and equals zero otherwise.

Note that if you think of the qubit as a spin, then the states with spin up or down point in parallel but 'opposite' directions in real space but they are represented by two *orthogonal* vectors in the state space of the quantum spin. The

state space picture therefore looks similar to that of the two polarizations of a photon, which are also in real space orthogonal. Yet there remains an essential difference, the qubit is what we call a spinor while the photon is a real vector. Note also that there are many choices of frame possible, for example the states $|+\rangle$ and $|-\rangle$ also form an orthonormal frame, as is depicted as the 'green frame' in Figure II.1.9.

The linear superposition principle

The expression (II.1.2) is an expansion of the state vector $|\Psi\rangle$ in an orthonormal basis $\{|i\rangle\}$. This a general rule: for any state vector in any $D-$dimensional Hilbert-space and any choice of basis one may write:

$$|\Psi\rangle = \Sigma_i^D \, \alpha_i \, |i\rangle \, , \qquad \text{(II.1.8)}$$

where once more the α_i are the components of the state vector in that particular basis. This *linear superposition principle* is a general property and is a consequence of the fact that the Hilbert space of quantum states is a vector space. Any linear combination of state vectors is (after normalization) again a possible quantum state. It follows from there also that any state can be expanded in a complete set of basis vectors, a property we have used above.

We can now show what it means to say that a state vector $|\Psi\rangle$ has unit length by writing:

$$\langle\Psi|\Psi\rangle = \Sigma_{i,j}\alpha_j^*\alpha_i\langle j|i\rangle = \Sigma_i|\alpha_i|^2 = 1\,. \qquad \text{(II.1.9)}$$

With what I just said, you may get worried about the Hilbert space for a real particle, because already in one dimension the configuration space is a line, corresponding to a continuum of classically allowed positions. But how then can you ever build a vector space of that continuous collection of points? That space has to be *infinite*-dimensional for a start.

Yes indeed, but in fact this can be done in a rigorous way! Our mathematical friends have shown that the space of functions on configuration space of the system is exactly the infinite-dimensional (!) Hilbert space of the type one needs to describe a particle with. The particle states correspond to functions on the classical configuration space, and as you may have guessed these are the famous wave-functions quantum people always talk about, the functions we introduced in Chapter I.4. The functions have to satisfy the additional condition that their squares are normalizable, so that they can be interpreted as probability densities. We will explore quantum states for particles and fields in more detail in Chapter II.5.

Ultimate simplicity: a single state system?

Let us make a small detour and imagine for a moment that you were to ask the silly question about what the quantum theory would look like for a system that has only a single state. A particle that only can be in one point. Should we waste our time with such a thing, which seems worse than thinking about how many angels can dance on the point of a needle, as the great theologian Thomas Aquinas appears to have worried about in the 13th century.

The quantum formalism would then say that this pin-point particle has a one-dimensional Hilbert space, so there is only one complex state vector that has to be normalized to one. It would look like:

$$|\psi\rangle = \alpha|0\rangle \text{ with } |\alpha|^2 = 1 \Rightarrow \alpha = e^{i\theta}\,.$$

There is only one phase and that phase is an overall phase which is not observable, as it drops out of the only possible probability amplitude $\langle 0|0\rangle = 1$, and so that finishes off the subject.
Except if we allow ourselves a minute amount of

freedom, maybe then....

So, let us imagine that this single state system represents the ground state of some real physical medium, and furthermore that possible other states in that medium have much higher energy, unreachable for the system all by itself, after all where would the energy come from? And if it were to jump up spontaneously by some quantum magic, it would plunge down instantly anyway. So we have a one-state Hilbert space for this system that corresponds to its ground state.

Now the critical readers are supposed to scratch their head and ask whether it is permitted to have two chunks of that material, both in that same ground state, but of course each with its own 'unobservable' phase. And they ask me: Sir, are two unobservable phases not a bit too much of obscurity? After all, what does *overall phase* mean in this context? Aha! Your point is well-taken. Two chunks making one system have one overall phase, but that leaves us with exactly one *relative phase*. But what is that good for, I may ask you in return. The puzzling point is indeed that we have two exactly identical pieces of exactly the same material, and we know all there is to know about them. There is nothing we can learn about them by making more measurements.

Well, let us sit back for a moment, and try to imagine some classical situations that are vaguely similar. I have two big chunks of material and I only talk about one variable, say temperature. There happen to be no thermal fluctuations because the material has infinite thermal conductivity! What you suggest is that we put one chunk in the freezer, and the other we keep exactly at room temperature. Each chunk in its own habitat is boring and stupid and nothing happens. But imagine we bring them out in thermally isolated boxes and put them on the table, and then take away the isolation at two facing sides and move them quite close. Sure enough the temperature difference will have an effect and heat will start flowing from the hot chunk to the cold chunk. In spite of the gap in between, there will be a thermal flow which is caused by the temperature difference. After this poor classical analogue (poor, because the temperature (difference) is of course a directly measurable observable for the individual subsystems), we rush back to our quantum chunks each with their own quantum vacuum phase angle. What we did pick up is the idea that we should bring them close together and see what happens.

The Josephson junction

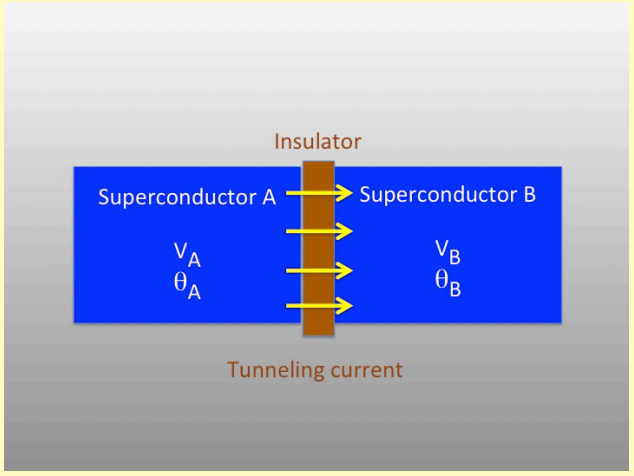
Often things don't have to be complicated to be interesting. What I am telling you is basically the story of the Josephson junction, referring to an effect that explains that having two slabs of superconducting material in the same superconducting ground state, but with different phase angle, one can indeed obtain a 'tunneling current' from one piece to the other! This is a truly remarkable physical effect, entirely due to the phase difference of two one-dimensional Hilbert spaces describing the same ground state. In spite of the fact that the slabs are not touching, they may quantum interact if you bring them close. And that quantum interaction turns the phase-difference into an observable.

So, how can we understand this more precisely using the Schrödinger equations for this system? We have two parts to the system with wavefunctions $|\psi_i\rangle \quad (i = A, B)$.

$$|\psi_i\rangle = e^{i\theta_i}|0\rangle .$$

The state is just the lowest state and is constant over the sample, and taking the inner product gives the Cooper pair density, the normalization is therefore that $\langle\psi_i|\psi_i\rangle = \langle 0|0\rangle = n$, because the phases

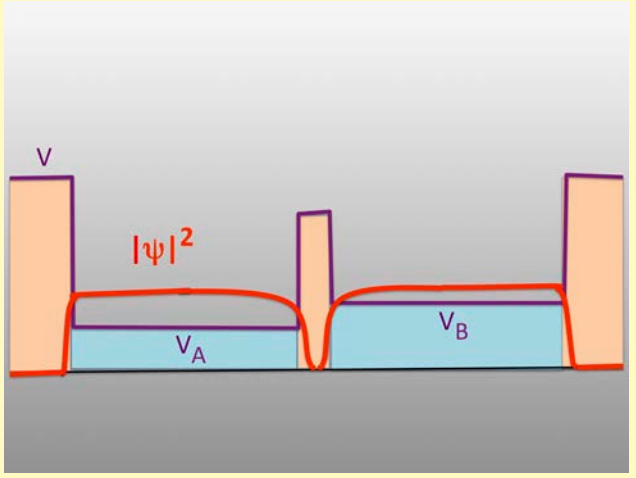
cancel. This state itself is a rather non-trivial affair but that doesn't concern us here. We just have a well-defined single state. If there is no coupling between the two pieces of super-conducting material, then this is the end of the story. The situation is completely static. We find ourselves talking chunks of superconducting material, in which nothing happens as long as you stay below the energy needed to break up a Cooper-pair.



(a): *Josephson junction.* Two 'identical' slabs of superconductor with an insulating layer in between. The ground states have few parameters, a homogeneous charge density $ne \simeq |\psi|^2$, a Potential V, and a phase angle θ.

Then life gets simple again, effectively it only has an angle which is hidden and does not really count as a degree of freedom. Trivial! So that's why we discuss it here as a case of ultimate simplicity, it really is less than a single particle, less even than a qubit!

But, imagine we bring the two pieces very close, then the wavefunction will decay exponentially outside the the space in between the two pieces, so once they are very close they can interact quantum mechanically, but not classically, the insulating material in between acts like a high potential barrier. Yet, the two pieces interact, which means that there is some weak coupling w. This situation is depicted in figure (b) and the interaction leads to



(b): *Charge density.* Potential landscape (purple curve) and charge density (red curve). The potential is minimal in the slabs, so the charges (Cooper pairs) are well confined. But at the boundaries of the slabs the wavefunction will decay exponentially, also on the insulator side, so if the insulator gap is narrow enough the wavefunctions of slab A and B will overlap and represent an interaction.

cross terms in the equations as follows:

$$i\hbar \frac{d|\psi_A\rangle}{dt} = eV_A|\psi_A\rangle + w|\psi_B\rangle\,, \qquad \text{(II.1.10)}$$

and a corresponding equation for $|\psi_B\rangle$ with the same V_B and a term $-w|\psi_A\rangle$.

A DC current. We start by considering the case with $V_A = V_B$. Now we don't have to solve the problem in all detail as we mainly want to know what the effect of the interaction on the charge densities is. We know that the electric current(density) J is defined as the time derivative of the charge (density), where the charge density is just $\rho = e\langle\psi_A|\psi_A\rangle$, with $-2e$ the charge of a Cooper-pair,

$$J = \frac{d\rho}{dt} = e\,\frac{d\langle\psi_A|\psi_A\rangle}{dt}\,. \qquad \text{(II.1.11)}$$

The right-hand side of this equation can be directly calculated from,

$$\frac{d\langle\psi_A|\psi_A\rangle}{dt} = \frac{d\langle\psi_A|}{dt}|\psi_A\rangle + \langle\psi_A|(\frac{d|\psi_A\rangle}{dt})\,.$$

After substituting the right-hand side of the equation (II.1.10) and its mirror we arrive at the following expression for the current:

$$J = \frac{-iew}{\hbar}(\langle\psi_A|\psi_B\rangle - \langle\psi_B|\psi_A\rangle) = \frac{2ewn}{\hbar}\,\sin(\theta_B - \theta_A)\,.$$

Defining the phase difference $\theta = \theta_B - \theta_A$, we obtain that

$$J = J_0 \sin\theta \quad \text{with} \quad J_0 = 2ewn/\hbar\,.$$

This is a stunning result! Apparently there is a DC current flowing through the junction without any potential difference, the current is basically driven by the phase difference between the two superconducting slabs!

An AC current. There is another important equation, which follows if we now in addition apply a voltage across the barrier. Then $V_A \neq V_B$, we can just solve for the phase difference θ to obtain

$$\frac{d\theta}{dt} \simeq \frac{2e}{\hbar}V\,, \qquad \text{(II.1.12)}$$

where V equals the potential difference $V \equiv V_B - V_A$. So we see that if we apply a voltage over the junction, the current becomes an AC current. This Josephson junction is a quantum device that has the remarkable feature that the frequency of the current measures the voltage!

The power delivered to the junction. Now the amount of energy is the power delivered to the junction over time, where the power is equal to the product of the current and the applied voltage JV. We can write this in terms of our fundamental angular variable:

$$U(\theta) = \int_0^t J\,V\,dt = \frac{J_0\hbar}{2e}\int_0^{\theta(t)} \sin\theta\,d\theta\,,$$
$$\rightarrow U(\theta) = \frac{J_0\hbar}{2e}(1-\cos\theta)\,. \qquad \text{(II.1.13)}$$

We find that this energy is periodic in the phase difference, which is not so surprising if you realize that the whole setup is periodic from the start. Yet, to get to a more complete understanding we should take another contribution to the energy into account.

The charging energy. You can think of this junction as a (super) capacitor, with two (super-)conducting plates and an insulator in between. We have an AC current $J(t)$ going through, so that a charge $Q(t)$ and a related voltage $V(t)$ will build up on the capacitor. The defining relation for the *capacity* C of the capacitor is $Q = CV$, and C is a property of the junction which does not depend on time.

There is a charging energy U_Q that builds up in the capacitor, which is given by the time integral,

$$U_Q = \int_0^{Q(t)} V\,dQ = \frac{1}{2C}Q^2\,. \qquad \text{(II.1.14)}$$

A mechanical analogue. Think of the total energy function as a Hamiltonian

$$H(Q,\theta) = \frac{1}{2C}Q^2 + \frac{J_o\hbar}{2e}[\cos\theta - 1]\,.$$

with of course also the relation,

$$Q = CV = \frac{\hbar C}{2e}\frac{d\theta}{dt}\,.$$

This reminds us of a simple particle Hamiltonian where the first term is the like the kinetic energy proportional to the momentum squared (the velocity being $d\theta/dt$), and the second like a potential energy. It describes a particle running around on the unit circle with an (angular) momentum Q proportional with the angular velocity $d\theta/dt$ in a nice periodic potential $U(\theta)$. This particle has a mass proportional to C and the strength of the potential is proportional to J_0. One can now check with the material we discussed in Chapter I.1, with $p = -Q$ and $q = \theta 2e/\hbar$, that (i) the dynamical equations correspond with the equations we derived for J and V, and (ii) that the total energy is indeed conserved for this mechanical system.
So this, in essence, basic quantum system, could in the end be mapped to a familiar classical system, where one can effectively apply one's good old Newtonian mechanics skills and intuitions.

This closes our Josephson-junction detour. Now you can appreciate the remarks we made in Chapter I.2 on units, equation (II.1.12) displays a direct relation between a frequency and the ratio of two universal constants which is by the way the fundamental unit of magnetic flux, $\Phi_0 = h/2e$. You can use this relation to measure that ratio, but also the other way around, knowing that ratio you can measure voltages extremely accurately. Indeed, this Josephson junction has many generalizations called Josephson's effects with ample applications.

This answers the so-called 'silly' question we started off with. The answer is that by introducing the interaction between two 'trivial' systems they become one, and there is only one unobservable phase left, while the other, relative, phase becomes a dynamical variable and acquires physical meaning of the utmost importance.

This intermezzo illustrates in my opinion something interesting about doing physics: it is not always a matter of taking as much as possible into account, but rather, it is rather that after stripping the problem back to its minimal form that essential insights are obtained. In other words, my advice to the alert reader would be: keep pestering your teachers with silly questions, because as you see, they may not be so silly after all and may lead to stunning answers!

By the way, it was the the Welshman Brian Josephson, who won the physics Nobel prize in 1973 at 33 years of age for the discovery of what is now called the *Josephson junction*, which is in essence the system we just described. He did the work in Cambridge as a student at the age of 22. In other words, we are never too old to learn and never too young to make a difference! We will come back into more detail to these matters in Chapter III.3 on condensed matter physics in Volume III of the book.■ □

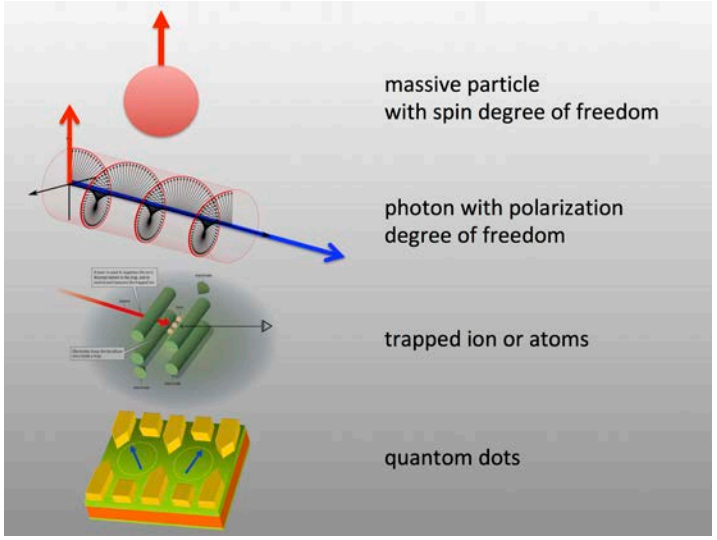


Figure II.1.10: *Qubit realizations.* Four possible qubit realizations: (i) an atom or particle that carries spin one-half like the electron, (ii) the photon, (iii) particles trapped in optical lattices having two well-separated levels, and (iv) spins in quantum dots.

Qubit realizations

Any well-defined two level quantum system can be thought of as representing a qubit. This could also mean we restrict ourselves to a subset of two specific states of a more elaborate quantum system. Examples of two state quantum systems are:
(i) *a particle that carries half a unit of spin* like the electron, the proton or neutron. These possess two basic spin states. If we measure its spin along any direction, we always find either spin 'up' or 'down'. This spin-$1/2$ property basically has no classical analog; we have introduced its discovery and its meaning on page 161 of Volume I.
(ii) *a photon* with a fixed frequency, which possesses two basic polarization states. The photon can oscillate in any direction perpendicular to its direction of motion, and as the photon necessarily moves with the velocity of light and just cannot be put to rest, this frame is always well defined. The polarization state can always be decomposed into two perpendicular basis states, say 'horizontal' and 'vertical'. We can arbitrarily designate one quantum state as 'spin up', represented by the symbol $|+1\rangle$, and the other 'spin down', represented by the symbol $|-1\rangle$. We illustrated some typical polarization states of a photon in Figure II.1.11. If both components are in phase with each other, we say that the photon is linearly polarized. If they are out of phase we speak of circular or elliptically polarized light, where we distinguish 'left-handed' or 'right-handed' polarization.
A photon is a qubit that necessarily travels with the speed of light. If we generate an electromagnetic wave, what we really do is making a beam of photons, and depending on the type of source, this beam maybe polarized or unpolarized. But if we make an ultra-short light pulse, it is possible to only produce a single photon.
(iii) *A particle (say atom or molecule) in one of two lowest energy states* which are well separated from the rest of the spectrum of states. A well-known example is the trapping of ions in an optical lattice.
(iv) In *quantum dots* it is possible to individually manipulate spin carrying degrees of freedom such as polarized electrons, and therefore these can in principle be assembled into quantum information processing devices.

Entanglement

It is in multi-particle and multi-qubit states that some of the most counter-intuitive and powerful aspects of quantum theory surface: in particular the notion of entanglement. In Figure II.1.12 we give a 'state of the union', a schematic overview of the multi-qubit type of states and how they are related. This schematic summarizes the content of this section.

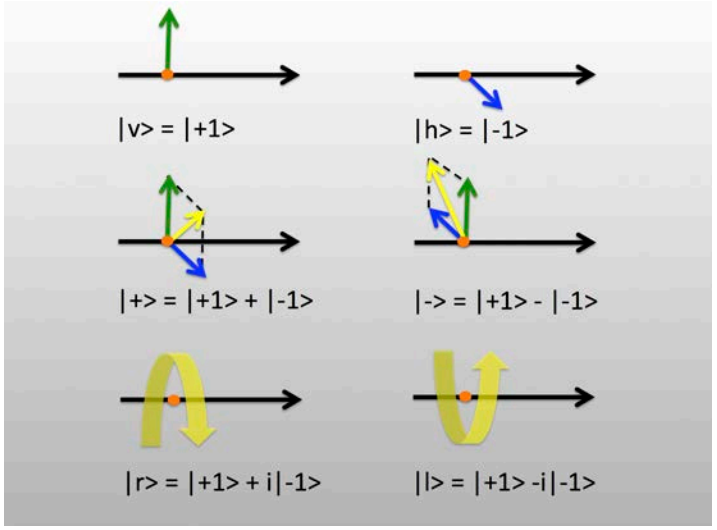


Figure II.1.11: *Photon polarizations.* Polarization states of the photon decomposed into the standard basis vectors $|+1>$ and $|-1>$. The top four are linearly polarized states, while the bottom two are circularly polarized. The polarizations in the three lines correspond to the eigenstates of the basic qubit observables, Z, X, and Y which will be defined after equation (II.2.2).

Multi-qubit states

A quantum computer needs systems of multiple qubits, called *quantum registers*. You may think of an array or network of n particles, each with its own spin. (As stated before, the formalism does not depend on the precise implementation, and it is possible to have examples in which the individual qubits correspond to degrees of freedom other than spin). The quantessence doesn't talk about how the qubits are realized, but about their underlying structural properties. The mathematical space in which the n qubits live is the *tensor product* of the individual qubit spaces, which we write as $\mathbf{C}^2 \otimes \mathbf{C}^2 \otimes ... \otimes \mathbf{C}^2 = \mathbf{C}^{2^n}$. For example, the Hilbert space for two qubits is $\mathbf{C}^2 \otimes \mathbf{C}^2$. This is a four-dimensional complex vector space spanned by the vectors $|1\rangle \otimes |1\rangle$, $|-1\rangle \otimes |1\rangle$, $|1\rangle \otimes |-1\rangle$, and $|-1\rangle \otimes |-1\rangle$. So tensor products are not about multiplying numbers or functions, but about multiplying spaces, where the product refers to the dimensions: the product of an m-dimensional and a n-dimensional space gives an $(m \times n)$-dimensional space. So multi-qubit states live in an exponentially larger state space ($d = 2^n$). For convenience we will often abbreviate the tensor product by omitting the tensor product symbols, or by simply listing the spins. For example

$$|1\rangle \otimes |-1\rangle \equiv |1\rangle|-1\rangle \equiv |1,-1\rangle\,.$$

The tensor product of two qubit states with state vectors $|\psi\rangle = \alpha|1\rangle + \beta|-1\rangle$ and $|\phi\rangle = \gamma|1\rangle + \delta|-1\rangle$ is the state

$$|\psi\rangle \otimes |\phi\rangle \equiv |\psi\rangle|\phi\rangle = \\ = \alpha\gamma|1,1\rangle + \alpha\delta|1,-1\rangle + \beta\gamma|-1,1\rangle + \beta\delta|-1,-1\rangle\,.$$

An basic feature of the tensor product is that it is distributive, i.e. $(\gamma|1\rangle+\delta|-1\rangle)\otimes|\psi\rangle = \gamma|1\rangle\otimes|\psi\rangle+\delta|-1\rangle\otimes|\psi\rangle$. We emphasize once more that whereas the classical $n-$bit system has 2^n states, the $n-$qubit system corresponds to a vector of unit length in a 2^n-dimensional complex space. It is a continuous space in fact a complex hypersphere. For example a three-qubit can be expanded as:

$$\begin{aligned} |\psi\rangle &= \alpha_1|1,1,1\rangle + \alpha_2|1,1,-1\rangle + \alpha_3|1,-1,1\rangle \\ &+ \alpha_4|-1,1,1\rangle + \alpha_5|1,-1,-1\rangle + \alpha_6|-1,1,-1\rangle \\ &+ \alpha_7|-1,-1,1\rangle\alpha_8|-1,-1,-1\rangle\,. \end{aligned}$$

As before it is convenient to denote the state vector by the column vector of its complex components $\alpha_1, \alpha_2, ..., \alpha_{2^n}$.

When dealing with multi-qubit states, we have to make clear distinctions between various types of states. These are important in discussions to come later, yet I want to present them here all at once, without elaborating too much on their specific roles yet. It is nice to compare them and contrast them with each other. First of all there are the so-called *pure states* and those are the states we have been talking about so far. The pure multi-qubit or multi-particle states break up into two types, the *separable* and the *entangled* states. The notion of entanglement and its dramatic physical implications are the subject of Chapter II.4,

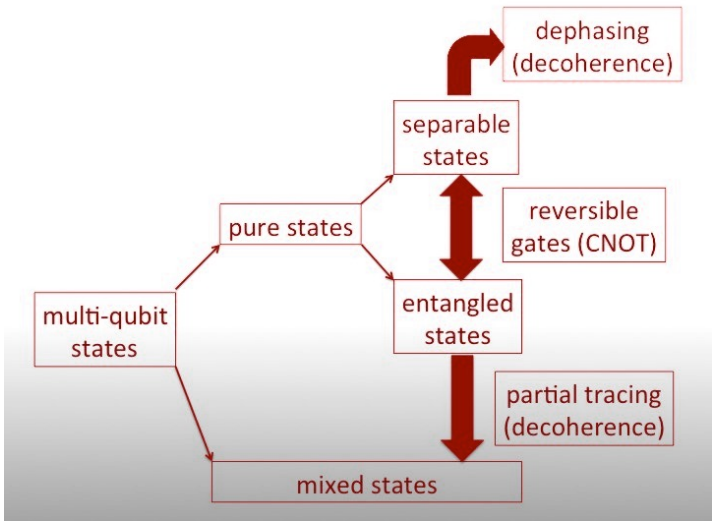


Figure II.1.12: *Multi-qubit states.* An overview of the types of multi-qubit states and the relations between them.

about the Einstein–Podolski–Rosen paradox and quantum teleportation.

If we talk about realistic quantum systems that couple to some environment or 'classical' measurement device, we often have to deal with states that are not pure but *mixed* states. To deal with both pure and mixed states it is convenient to introduce the *density operator*, which provides a unified framework for all types of states. This concept was introduced by Von Neumann in the early days of quantum theory as an alternative to the wavefunction or state vector approach. These are the topics that I will focus on in the remainder of this section.

Entangled states

> When two systems, of which we know the states by their respective representatives, enter into temporary physical interaction due to known forces between them, and when after a time of mutual influence the systems separate again, they can no longer be described in the same way as before, viz. by endowing each of them with a representative of its own.
> I would not call that *one* but rather *the* characteristic trait of quantum mechanics, the one that enforces its entire departure from classical lines of thought. By the interaction the two representatives [the quantum states] have become *entangled*.
>
> *E. Schrödinger, 1935*

Entanglement is a direct consequence of the linear superposition principle applied to multi-qubit or multi-particle states. If qubits are entangled this means that successive measurement outcomes on the two qubits will be highly correlated, implying that quantum theory is fundamentally non-local.

The quantum states of systems consisting of spatially separated components (e.g. two particles) can be *entangled*, which implies that they can no longer be treated independently and therefore measurements made on one can have instantly consequences for the other! This is indeed a quantessential feature of reality that dramatically departs from the classical description of such a system. It is this 'entanglement' property that lies at the root of a zoo of so-called quantum paradoxes, such as Schrödinger's cat and the EPR paradox and more generally the quantum measurement problem. But it is also essential for understanding the Bell inequalities which pose a rigorous quantitative bound on classically allowed correlations; bounds that have been observed to be violated in quantum systems. Entanglement furthermore plays an essential role in fashionable and promising subjects like quantum teleportation. We return to these topics in Chapter II.4. In this section we will merely touch on some of these aspects.

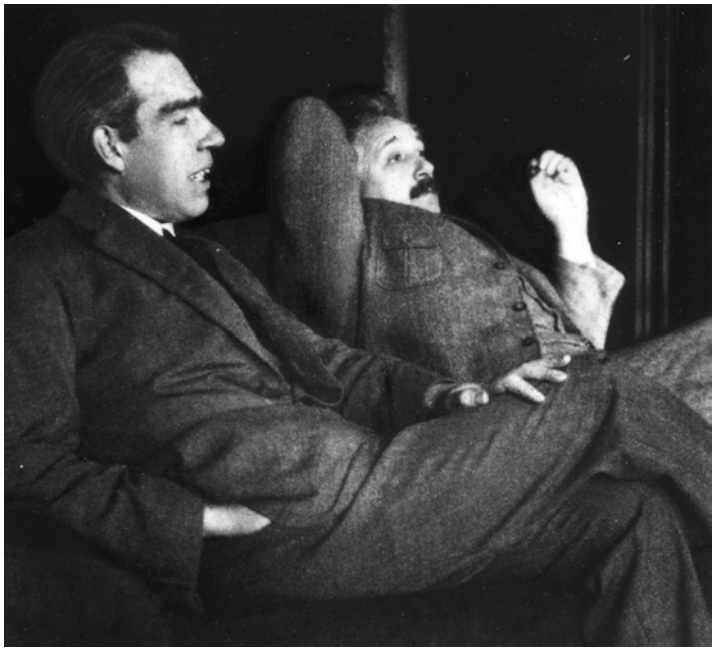


Figure II.1.13: *Bohr and Einstein in one of their debates.* (Source: (Photo made in 1930 by Paul Ehrenfest, courtesy AIP Emilio Segrè Visual Archives)

Schrödinger's cat

When we have more than one qubit an important practical question is when and how measurements of a given qubit depend on measurements of other qubits. Because of the deep properties of quantum mechanics, qubits can be coupled in subtle ways that produce consequences for measurement that crucially differ from classical bits. Understanding this has proved to be important for questions relating to quantum computation and information transmission. To explain this we need to introduce the opposing concepts of separability and entanglement, which describe whether measurements on different qubits are statistically independent or statistically dependent.

This notion of entanglement as a necessary consequence of the quantum postulates led to the infamous problem of Schrödinger's cat. This problem was well described by Schrödinger himself:[3]

> '[...]Man kann auch ganz burleske Fälle konstruieren. Eine Katze wird in eine Stahlkammer gesperrt, zusammen mit folgender Höllenmaschine (die man gegen den direkten Zugriff der Katze sichern muss): in einem Geigerschen Zählrohr befindet sich eine winzige Menge radioaktiver Substanz, so wenig, dass im Laufe einer Stunde vielleicht eines von den Atomen zerfällt, ebenso wahrscheinlich aber auch keines; geschieht es, so spricht das Zählrohr an und betätigt über ein Relais ein Hämmerchen, dass ein Kölbchen mit Blausäure zertrümmert. [...]
> '

and[4]

> [...] Hat man dieses ganze System eine Stunde lang sich selbst überlassen, so wird man sich sagen, dass die Katze noch lebt, wenn inzwischen kein Atom zerfallen ist. Der erste Atomzerfall würde sie vergiftet haben. Die Psi-Funktion des ganzen Systems würde dass so zum Ausdruck bringen, dass in ihr die lebende und die tote Katze

[3] It is also possible to construct very burlesque fables. A cat is locked into a steel chamber, together with a poisoning contraption consisting of a hammer and a flask (which must be secured against direct access by the cat): and a Geiger counting tube containing a minute amount of radioactive substance, so little that in the course of an hour perhaps one of the atoms breaks up, but equally probably none; if it happens, then the counting tube responds and, via a relay, releases the hammer that crushes a little flask with blue-acid.

[4] [...] After one has left this whole system for an hour, one will say that the cat is still alive if no atom has decayed, as the first atomic decomposition would have poisoned it. The wavefunction of the whole system would thus express the fact that in it the living and the dead cat are mixed or smeared in equal parts. That an indeterminacy confined to the atomic realm translates into indiscernible indeterminacy, which can then be removed by direct observation. This prevents us, in such a naive way, from considering such a 'washed out model' as an image of reality ...

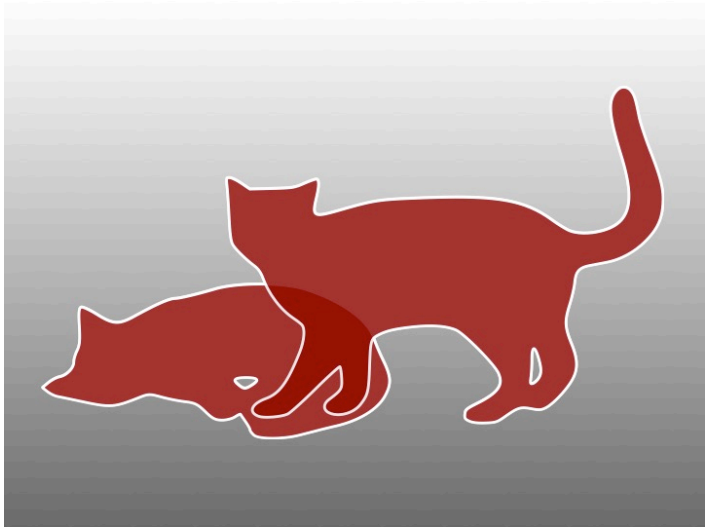


Figure II.1.14: *Schrödinger's cat state.* Artist impression of a quantum cat in the state: $|\psi_{\rm cat}\rangle = |{\rm alive}\rangle + |{\rm dead}\rangle$. (Source: JSTOR Daily.)

> (s. v. v.) zu gleichen Teilen gemischt oder verschmiert sind. Das Typische an solchen Fällen ist, dass eine ursprünglich auf den Atombereich beschränkte Unbestimmtheit sich in grobsinnliche Unbestimmtheit umsetzt, die sich dann durch direkte Beobachtung entscheiden lässt. Das hindert uns, in so naiver Weise ein "verwaschenes Modell" als Abbild der Wirklichkeit gelten zu lassen...'

A cat in our classical world can either be dead or alive, and taking this quantum assumption to its logical extreme, this cat could in principle be in a state that it is a linear superposition of 'alive' and 'dead'. This property of quantum mechanics is simple to spell out but is radically different from the way we talk about physical states in classical physics. This difference derives directly from the quantessential principle that allows us to consider linear superpositions of states, which therefore seems problematic from the start.

The cat sits in a closed box with some food but also with a lethal contraption consisting of a small quantity of a radioactive substance, or a single metastable atom for that matter. If that atom decays, it emits a photon that hits a detector which subsequently triggers a device which breaks a little capsule filled with a poisonous gas that in turn will kill the cat. This unfortunate scenario suggests that in this situation the states of the atom labeled $|{\rm decayed}\rangle$ or $|{\rm not\ decayed}\rangle$ are entangled with the states $|{\rm alive}\rangle$ or $|{\rm dead}\rangle$ and we write:

$$|\psi_{\rm cat}\rangle = |{\rm not\ decayed}\rangle \otimes |{\rm alive}\rangle + |{\rm decayed}\rangle \otimes |{\rm dead}\rangle\,,$$

because the other states in the ${\rm atom} \otimes {\rm cat}$ state space have zero coefficient, and we have assumed that both terms are equally probable. What the formula above expresses is that the undetermined state of the atom is entangled with the states of the cat.

It seems a far-out proposal of a fundamental theory of nature to take such states seriously. At the heart of this problem lies the following question: if quantum mechanics is the underlying reality of everyday life described by the laws of classical physics, then it should be possible to understand these classical laws from the quantum laws. There may be no logic that leads you from classical to quantum theory but there should be a derivation of the laws of classical physics starting from the quantum laws, because classical physics is just an approximation of quantum physics, and such approximations should be well understood. We should compare this to how classical Newtonian mechanics can be obtained as the limit of relativity where we send the speed of light c to infinity. The analogy suggests that in quantum theory we just have to send Planck's constant $\hbar$ to zero, and yes, in many cases this is what we have to do, but such direct approaches do not resolve issues like Schrödinger's cat. The question has lead to numerous deep philosophical arguments among physicists and philosophers right from the inception of quantum mechanics in the beginning of the twentieth century, and we will return to 'Schrödinger's cat' in later chapters. For the moment we just want to give a more accurate definition of

the different types of states for simpler systems.

Entangled vs separable states

Let us now turn to precise definitions of multi-qubit states. The n-qubit state is called *separable* if it can be written as a single product of n single-qubit states[5], i.e. if it can be written as $n-1$ tensor products of sums of qubits, with each factor depending only on a single qubit. An example of a separable two-qubit state is

$$|\psi\rangle = \frac{1}{2}(|1,1\rangle + |1,-1\rangle + |-1,1\rangle + |-1,-1\rangle)\,,$$

because it can be written like

$$|\psi\rangle = \frac{1}{2}(|1\rangle + |-1\rangle) \otimes (|1\rangle + |-1\rangle) = |+\rangle\,|+\rangle\,.$$

If an n-qubit state is separable, then measurements of individual qubits are statistically independent, i.e. the probability of outcomes of a series of measurements of different qubits can be written as a product of probabilities of the measurements for each qubit. These outcomes are uncorrelated and the overall outcome is therefore independent of the order in which these measurements are performed.

If an n-qubit state is not separable, then it is per definition *entangled*. An example of an entangled two-qubit state is,

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|1,1\rangle + |-1,-1\rangle)\,, \qquad \text{(II.1.15)}$$

which indeed is a linear superposition which cannot be factored into a single product. If we have a pair of qubits in an entangled state, subsequent measurements of the individual qubits do depend on each other. If you first make a measurement on the first bit, then that measurement will *instantaneously* affect the two-bit state and possibly the state of the other bit, even if that is spatially arbitrarily far away. The measurement thereby influences a later measurement outcome of the second bit. Now the use of that word 'instantaneous' should make you feel uneasy in view of the theory of relativity, and correctly so. Some great physicists – like Einstein to mention one – felt the same way and preceded you. This thought-provoking question unleashed a deep, but also longwinded debate about the foundations of quantum theory, already among its founders.

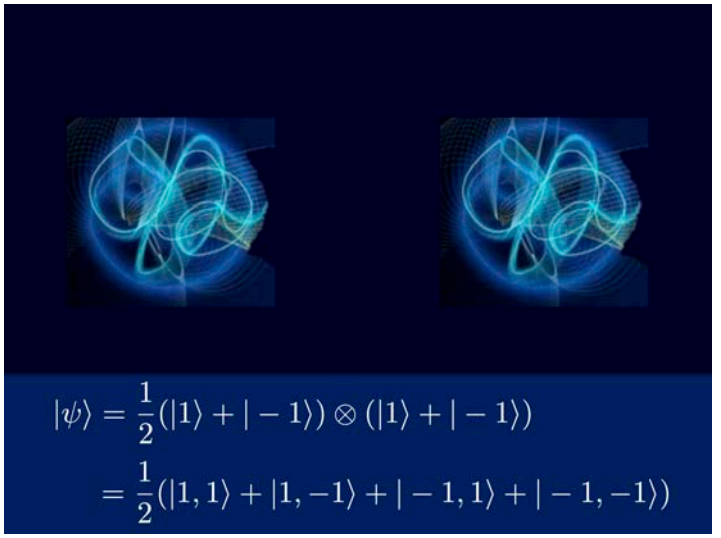


Figure II.1.15: *Separated pair.* The state vector for the pair is a product of the individual state vectors.

Let me illustrate this point for the examples we gave above. Suppose we do an experiment in which we measure the spin of the first qubit and subsequently measure the spin of the second qubit. For both the separable and entangled examples, there is a 50% chance of observing either spin up or spin down on the first measurement. Suppose it gives spin up. For the separable state this transforms the

[5] Strictly speaking this is only true for pure states, which we define in the next section.

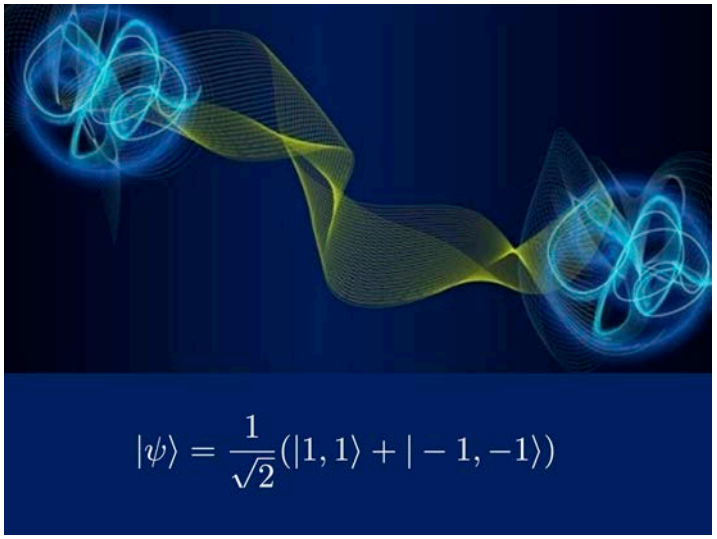


Figure II.1.16: *Entangled pair.* Artist impression of what an entangled state of a pair would look like. It gives at least a feeling for it, maybe more so than the formula below telling you exactly what it is.

state vector as follows:

$$\frac{1}{2}(|1\rangle + |-1\rangle) \otimes (|1\rangle + |-1\rangle)$$
$$\rightarrow \frac{1}{\sqrt{2}}|1\rangle \otimes (|1\rangle + |-1\rangle) = \frac{1}{\sqrt{2}}(|1,1\rangle + |1,-1\rangle)\,,$$

it means that the measurement projects the initial qubit state $|\psi\rangle$ on the first line onto the state $|\psi'\rangle$ in the second line. One may verify that the probability amplitude in analogy with equation (II.1.5) equals $|\langle\psi'|\psi\rangle|^2 = 1/2$ as it should. The same probability would have resulted for the spin down measurement.[6]

So only the $|1\rangle$ component of the first qubit survives after the measurement. If we now measure the spin of the second qubit in the state $|\psi'\rangle$, the probability of measuring spin up or spin down is still 50%. And as mentioned before, the previous measurement on the first qubit has no effect on the second measurement. As we have already noted, it is a generic property of separable states that subsequent measurement outcomes on individual spin states are independent, and the outcomes do not depend on the order in which we perform the measurements.

Let us now consider a similar experiment on the entangled state of equation (II.1.15) and observe spin up in the first measurement. This changes the state-vector to

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|1,1\rangle + |-1,-1\rangle) \rightarrow |\psi'\rangle = |1,1\rangle\,. \quad \text{(II.1.16)}$$

(Note the 'disappearance' of the factor $1/\sqrt{2}$ due to the necessity that the projected state vector remains normalized). If we now measure the spin of the second qubit, we are certain to observe spin up! Similarly, if we observe spin down in the first measurement, we will also see that in the second qubit with 100% certainty. For this entangled example the measurement outcomes are completely correlated – the outcome of the first completely determines the second, and the state is therefore called maximally entangled. As this also holds for entangled qubits which are light years apart, this instantaneous effect on the state implies a puzzling if not bizarre form of non-locality in the quantum world that at first sight appears to violate causality.

Bertlmann's socks. There has been a debate among physicists like John Bell and others about what it is that sets quantum entanglement really apart. The conundrum goes by the name *Bertlmann's socks*. Mr Bertlmann, a real-life early collaborator of Bell at CERN, happens to always wear socks of a different color. So, Mr Bertlmann, whose socks have risen to eternal fame, constitutes a system which has the unusual property that if you get to see one of his socks to be 'red' for example, then instantaneously you are able to conclude that the other sock has the property 'not red'. So here is a form of non-locality. You measure one sock and are hundred percent sure about a property of the other sock which is elsewhere. So, the conditional probability given sock #1 is red, for sock #2 to be

[6] We will deal with the observables and measurements more extensively in the next chapter.

'not red', is one. Is that not a classical version of quantum entanglement? It looks like it, the states of the socks are highly correlated indeed, and knowing the state of the first affects the probability distribution for the other. It doesn't change the socks or their color: it just affects the probability of measurement outcomes. And there is nothing unusual, absurd or stunning about that. It is very much true that the state of the socks is not affected. There is no signal exchanged between the socks, since they are in a definite state which is there to stay.

The quantessence of entanglement. The quantum catch is that there is one additional feature in the quantum framework that has no classical analogue and which sets the EPR paradox apart. In the qubit experiment there is an additional freedom for making the measurements, one is free to choose the frame or polarization of a measurement. In the example of the entangled state given in (II.1.16), we could have chosen the measurement for the first qubit not in the $(|1\rangle, |-1\rangle)$ frame, but for example of in the $(|+\rangle, |-\rangle)$ frame. Then, given the outcome of that measurement for example to be plus one, we know that the second qubit has to be in the $|+\rangle$ state. Keeping the measurement for the second qubit as before in the $(|1\rangle, |-1\rangle)$ frame, the probability to find the outcome to be plus or minus one is 50% for each. This dependence of the probability of the second outcome on the choice one makes for the first measurement is what makes the situation non-local, because now, dependent on the frame choice and the outcome plus one for the first measurement, the second qubit flips instantaneously to the $|1\rangle$ or $|+\rangle$ state. And this looks very much like an instantaneous action at a distance, the state of the second *is* affected, and therefore causality should be violated. Is it?

The answer is: no! As we have already, and will explore more extensively in the following chapters, the quantum state is like a probability amplitude, which encodes a probability distribution for measurement outcomes. Multi-qubit states, separable or entangled, encode all possible correlations that may or may not exist between sequences of measurement outcomes. And a closer look at the examples given above does precisely that, they show how unconditional probabilities, turn into conditional probabilities which are different indeed. And since in the quantum world there are basically only probabilities, the measurements of entangled states are easier to grasp if you think of 'states' as encoding probability distributions. We return to these questions in the section on the Einstein–Podolsky–Rosen paradox in chapter II.4.

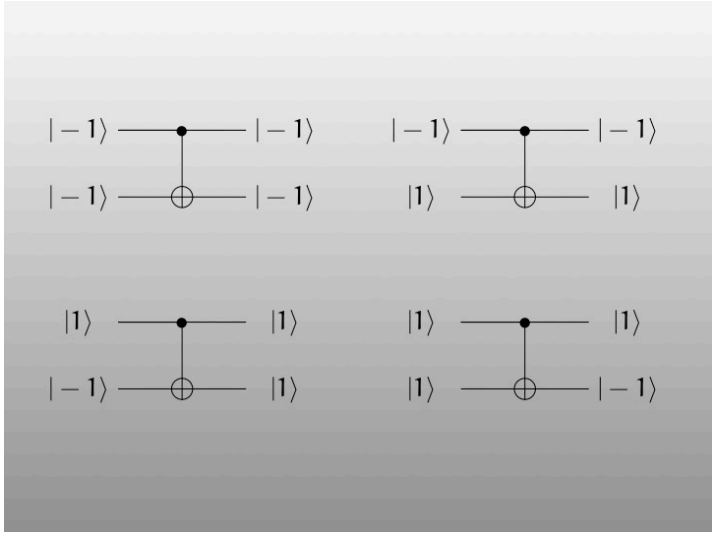


Figure II.1.17: *CNOT gate.* The circuit diagram is basically a two-qubit interaction diagram, representing the action of the CNOT gate on the four possible two-qubit basis states. As pointed out in Figure II.1.5, the dot on the upper qubit denotes the control and the cross is the symbol for the conditional one-qubit NOT gate.

From separable to entangled and back

For two qubits in a separable state to get entangled they need to interact somehow. In quantum information language that would mean that they have to be acted upon by some two-qubit gate. Let us take our favorite CNOT-

gate of Figure II.1.5, it acts on the state $|A\rangle \otimes |B\rangle$ as follows:

$$\mathrm{CNOT} : |A\rangle \otimes |B\rangle \Rightarrow |A\rangle \otimes |[-AB]\rangle \,.$$

In other words, the CNOT gate flips the state of B if $A = 1$, and does nothing if $A = -1$.

For convenience we give the explicit action on the basis states in Figure II.1.17, which allows you to verify that if we let the CNOT gate act on the separable state $|+\rangle \otimes |-1\rangle$ it indeed generates a maximally entangled state (II.1.15):

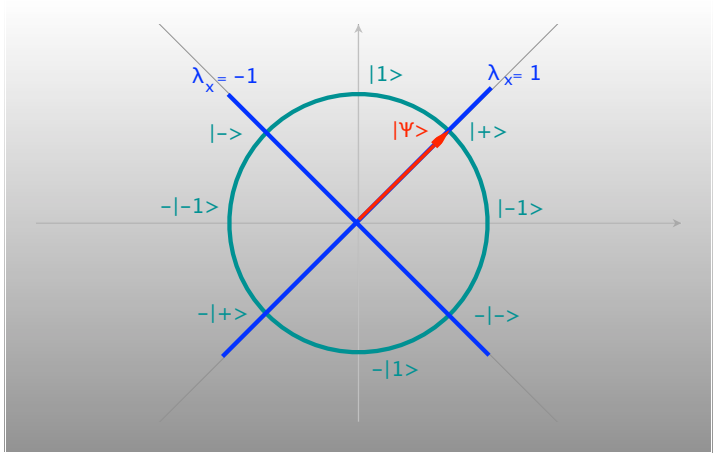
$$\mathrm{CNOT} : |+\rangle \otimes |-1\rangle \Rightarrow \frac{1}{\sqrt{2}}(|1,1\rangle + |-1,-1\rangle) \,. \quad \text{(II.1.17)}$$

Note that this gate is reversible, as one can immediately see from the figure, $\mathrm{CNOT}^2 = 1$.

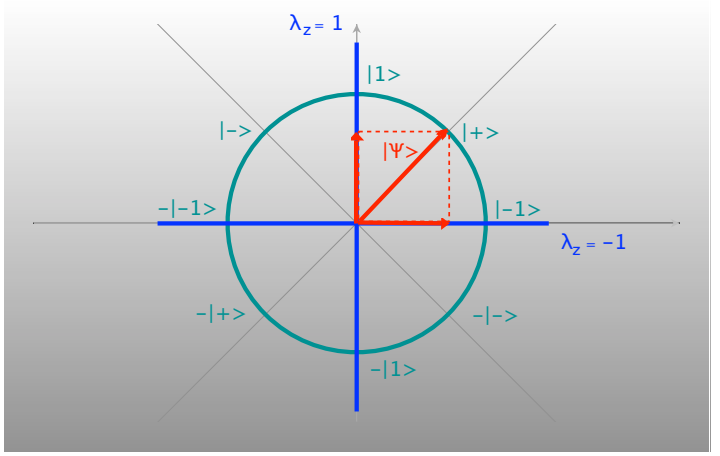
In fact, from an intuitive point of view the ability to generate substantial speed-ups using a quantum computer vs. a classical computer is related to the ability to operate on the high dimensional state space including the entangled states. To describe a separable n-qubit state with k bits of accuracy we only need to describe each of the individual qubits separately, which only requires of the order of nk bits. In contrast, to describe an n-qubit entangled state we need of the order of k bits for each dimension in the Hilbert space, implying that we need of the order of $2^n\, k$ bits. If we were to simulate the evolution of an entangled state on a classical computer we would have to process all these bits of information and the computation would be extremely slow. Quantum computation, in contrast, acts on all this information at once – a quantum computation acting on an entangled state is just as fast as one acting on a separable state. This is exactly the type of parallelism at the intermediate stages of computing that we referred to before. Thus, if we can find situations where the evolution of an entangled state can be mapped into a hard mathematical problem, we can achieve spectacular speed-ups.

Mixed versus pure states

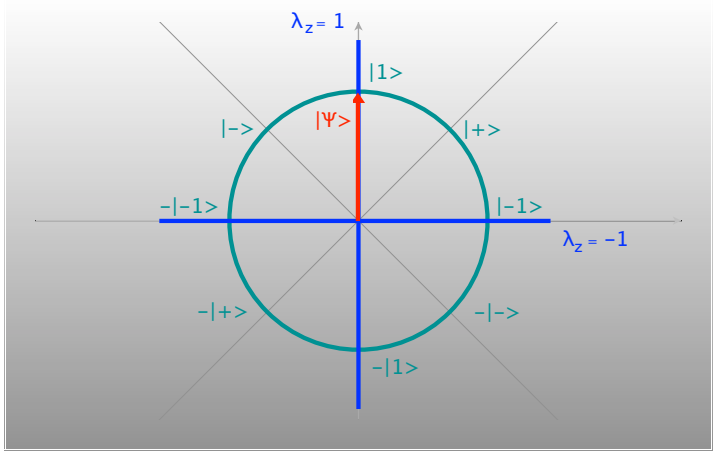
The states we have been dealing with so far were statistically pure, or more simply, *pure states*. In spite of the quantessential uncertainties in such states, the state vector is the most complete knowledge about a quantum state that is available. In real life however it may prove very difficult to prepare a system in a pure state. After all, quantum phenomena are not that easy to detect, which means that pure states apparently are not so common. Somehow a lot of the quantum stuff gets washed away in ordinary life, quantum does not hit the eye, so to speak. The reason is that quantum systems are permanently interacting with their environment, and it is only in situations where we take exceptional care to protect our quantum system from those influences, that we can observe pure quantum behavior. This is not easy; it certainly is not the case in most situations which arise naturally, and that is precisely why we perceive the world around us as completely classical. Turning this reasoning around you may ask that given the underlying world is entirely quantum, why there is such a thing as the classical world, and how it comes about. How can we understand the laundering of all that quantum exotica? This is the basic question one has to face in a detailed treatment of quantum measurement, which has to account for how we can start up with a quantum process and end up reading dials and counting macroscopic signals like clicks, or pulses and what not. It is here that we have to introduce the concept of a *mixed state*, and contrast it with a pure state whether entangled or not. And indeed it is often through the interaction with the environment that states get 'mixed up', just like humans do. We have to deal with mixed up people all the time, and we learned to deal with that! Let us be pedantic and illustrate the distinction between a pure and a proper mixed state with the experimental setup depicted in Figure II.1.18. An incoming beam is polarized and each of the particles is in the pure state $|+\rangle$ i.e. with $X \sim +1$. Now we send them through a Z polarizer in (b). What we find behind the polar-



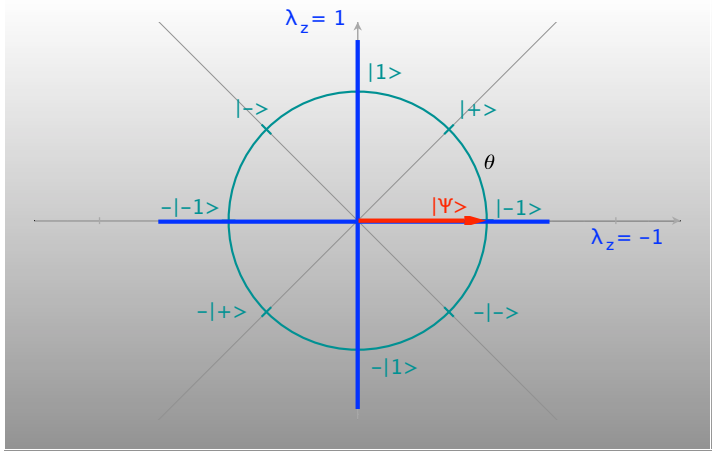
(a) All incoming particles are in the pure state $|\psi\rangle = |+\rangle$.



(b) The incoming beam goes through a Z polarizer



(c) Half of the particles in the outgoing beam are in the pure state $|\psi\rangle = |1\rangle$.



(d) The other half of the particles in the outgoing beam are in the pure state $|\psi\rangle = |-1\rangle$.

Figure II.1.18: *Mixed state.* Graphical representation of how to prepare a beam of particles in a proper mixture and corresponding mixed state. For the incoming particles in figure (a) the density matrix is given by equation (II.1.23), in the outgoing beam (c)+(d) the particles have the density matrix of equation (II.1.24)

izer is a beam of particles, but now 50% of the particles is in a pure state $|1\rangle$ (c) and the other 50% is in a pure state $|-1\rangle$ (d). Now we could combine the particles in single beam (without letting them interfere), then each of the particles is still in a pure state, but the beam is now a statistical mixture of particles in $|\pm 1\rangle$ states. The beam represents a classical ensemble of particles that is in a *mixed state*. This is called a *proper mixture* to be distinguished from the improper mixture to be discussed shortly. In the present situation there is a non-zero probability for the particle to be in any one of two pure states. So picking out one particle there is a 50% chance it is in a pure up-state and a 50% chance it is in the down-state, but – and this is crucial – it is not a state corresponding to a linear superposition of the up and down-state, that would just be the 'plus' state, $|+\rangle$!

This mixed state is a classical statistical mixture and not a quantum superposition. What makes this setting a bit confusing is that we now have two types of probability to keep track of, the quantum probabilities we have been talking about so far and in addition the probability distribution of the classical ensemble.

To further clarify this notion, let me point out that a naive but *wrong* way is to write for the state of the particle something like $|\psi_{mix}\rangle = \sum_i p_i|\psi_i>$. This looks dangerously close to the usual expansion of a pure state into a certain basis $|\psi\rangle = \sum_k \alpha_k|\psi_k>$. But there the coefficients are probability amplitudes α_k leading to probabilities $p_k \simeq |\alpha_k|^2$. To put it differently, with the troublesome trial notation I just proposed we would get that the expectation or average value[7] of an observable A in a mixed state $|\psi_{mix}\rangle$ would become $\langle\psi_{mix}|\, A\, |\psi_{mix}\rangle \sim \sum p_i p_j \ldots$, an expression that is proportional to probabilities squared, which makes no sense.

[7] I apologize for getting ahead of myself, as observables and their expectation values are to be discussed in detail in the next chapter on page 285.

What we want is a weighted average over ordinary pure state expectation values:

$$\langle A\rangle = \sum_a p_a \langle\psi_a|A|\psi_a\rangle\,. \qquad \text{(II.1.18)}$$

In this expansion the states $|\psi_a\rangle$ are some set of pure states. These don't have to be orthogonal, so it could be that $|\psi_1\rangle = |1\rangle$ and $|\psi_2\rangle = |+\rangle$ for example. It is for this reason that once we admit both mixed states and pure states it is almost imperative to use the density matrix formalism because it treats both type of states on an equal footing.

The density operator

The famous mathematical physicist John von Neumann developed an alternative formalism for quantum mechanics in terms of what is called a *density operator*, which basically replaces the wavefunction, or state vector, right from the start.

The density operator formulation, as we will see shortly, leads in a natural way to the definition of what is called the *Von Neumann entropy* for a quantum system.

Proper mixtures. Consider as we just did, a mixed state in which there is a probability p_a for the system to have wavefunction ψ_a and an observable A with an expectation value (II.1.18). The density operator defined for a pure state is just the projection operator we introduced in (II.1.6) for that state:

$$\rho = P_i = |\psi_i\rangle\langle\psi_i|\,,$$

and the density operator for a properly mixed state is quite naturally defined as:

$$\rho = \sum_i p_a\, |\psi_a\rangle\langle\psi_a|\,, \qquad \text{(II.1.19)}$$

which reduces naturally to the pure state case above if $p_a = p_i = 1$ for a single value of i. To obtain the *density*

matrix, we have to expand this operator in an orthonormal basis. We start with

$$\rho = \sum_{ajk} p_a \alpha_j^{(i)} \alpha_k^{*(i)} |\chi_j\rangle\langle\chi_k| \,,$$

where the $\alpha_j^{(i)}$ are the expansion coefficients of the pure state $|\psi_a\rangle$ in the $\{|\chi_j\rangle\}$ basis. The density matrix is defined by the matrix elements of the density operator:

$$\rho_{mn} = \langle\chi_m|\rho|\chi_n\rangle = \sum_i p_i \alpha_m^{(i)} \alpha_n^{*(i)} \qquad \text{(II.1.20)}$$

With density matrices the notion of a $trace$ is convenient. Recall that the trace of a matrix is the sum of the diagonal elements, so we have for example that,

$$\mathrm{tr}\,\rho = \sum_i p_i (\sum_m \alpha_m^{(i)} \alpha_m^{*(i)}) = \sum_i p_i = 1 \,,$$

because the sums over m equal one for each value of i. The expectation value (II.1.18) is compactly expressed as:

$$\langle A\rangle = \mathrm{tr}\,(A\rho) \,, \qquad \text{(II.1.21)}$$

and because the trace $\mathrm{tr}(A\rho)$ is independent of the chosen basis this expression can be evaluated in any convenient basis, and so provides an easy way to compute expectation values in any state. Note that for a pure state $p_a = 1$ for one particular value of $a = i^*$, and $p_a = 0$ for $a \neq i^*$. In this case the density matrix has rank one. This becomes clear if we write the matrix ρ in a basis in which it is diagonal, because then there will only be one non-zero element. When there is more than a single non-zero value of p_a it is a mixed state and the rank is larger than one.

Finally note that if we chose the unit matrix as the trivial observable we get the trace of ρ itself, which equals one by definition. This property will be used if we consider partial traces, which refer to the density matrix of subsystems, later on. The best way to think about the density matrix of a proper mixture is as a classical distribution over pure quantum states. It is an essential concept if we want to understand and describe quantum measurements in more detail. In particular if we are to include the measurement apparatus in the analysis, and want to understand how we get from quantum to classical physics: from a pure quantum state to a macroscopic pointer on a dial.

To get a better feeling for how a density matrix works, consider a few simple examples of single qubit states. First look at a spin in a pure state with $|\psi\rangle = |1\rangle$. The density operator corresponds to the corresponding projection operator.

$$\rho = |1\rangle\langle 1| \Leftrightarrow \rho_{mn} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} . \qquad \text{(II.1.22)}$$

The expectation of the spin polarization operator along the z-axis becomes

$$\mathrm{tr}(Z\rho) = \mathrm{tr}\left(\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\right) = \mathrm{tr}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = 1 \,,$$

as expected. Likewise we could construct the *density matrix* corresponding to another pure state $|+\rangle$ as

$$\rho = |+\rangle\langle +| = \frac{1}{2}(|1\rangle + |-1\rangle)(\langle 1| + \langle -1|) \Leftrightarrow \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} . \qquad \text{(II.1.23)}$$

Now the expectation value of Z is

$$\mathrm{tr}(Z\rho) = \frac{1}{2}\mathrm{tr}\left(\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right) = \frac{1}{2}\mathrm{tr}\begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix} = 0 \,,$$

as it should. If, however, the system is in a mixed state with 50% of the population spin up and 50% spin down the density matrix becomes

$$\rho = \frac{1}{2}\,(|1\rangle\langle 1| + |-1\rangle\langle -1|) \Leftrightarrow \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} . \qquad \text{(II.1.24)}$$

In this case the expectation of the spin along the z-axis, which is $\mathrm{tr}(Z\rho)$, is zero again as it should be, because the probability for a particle in the mixed state to contribute $+1$ is equal to the probability to contribute -1 to the expectation value. The particle represents a classical statistical ensemble of particles that are either in a definite quantum 'up' or a definite quantum 'down' state.

Quantum entropy

The introduction of the density matrix allowed Von Neumann to extend the fundamental concept of entropy to the quantum domain. He defined the entropy of a quantum state in analogy with the Gibbs entropy for a classical ensemble as

$$S(\rho) = -\text{tr}\ \rho \log \rho = -\sum_i p_i \log p_i\,. \qquad \text{(II.1.25)}$$

Where the right-hand side directly follows from the definition (II.1.19). *The entropy of a quantum state provides a quantitative measure of 'how mixed' a system is because the entropy of a pure state is equal to zero, whereas the entropy of a mixed state is always greater than zero.* Let us check this with the examples of the previous subsection. In the cases with pure states we have first considered (II.1.22) where $p(+1) = 1$ and $p(-1) = 0$ which for the quantum entropy yields $S_{\text{pure}} = -1 \log 1 - 0 \log 0 = -1 \cdot 0 = 0$. This reflects that if you know the pure state a system is in, you know everything there is to know about it, and therefore there is no hidden information and the entropy should be zero and happily that is true. For the properly mixed case of (II.1.24) we found $p(+1) = p(-1) = 1/2$, and we obtain that $S_{\text{mixed}} = -2 \cdot \frac{1}{2} \log \frac{1}{2} = \log 2$, which corresponds to the information of one bit. And it is here that we make contact with the definition by Shannon of information being proportional to the entropy as defined by Boltzmann and Gibbs. We see quite generally that a mixed state corresponding to an equal probability distribution over the pure states one has $\rho = \frac{1}{N} \sum |i\rangle\langle i|$, which will have the maximal entropy $S = \log N$ corresponding to the good old Boltzmann formula. All this underscores the remark that a (properly) mixed state is just a classical distribution over quantum states.

Entanglement entropy

We just saw how the Von Neuman entropy yields a quantitative measure of 'how mixed' a quantum state is. The entropy of a pure state (that may be entangled or not) is always equal to zero, whereas the entropy of a mixed state is always greater than zero. So, why inventing the term *entanglement entropy* if the entropy of an entangled state is always zero? The logic is somewhat oblique in that the term in fact refers to the entropy of a mixed state which is obtained after one traces out 'part' of the density matrix of an entangled state. For this reason such states are referred to as *improper mixtures*, in contrast to the *proper mixtures* which refer to the cases we discussed before where the state is a statistical mixture of pure states.

Partial traces and improper mixtures. In certain situations there is indeed a close relationship between entangled and mixed states, and that is what I would like to explain next. It entails a mechanism that plays a vital role in explaining the all-important fact that the world we perceive is classical rather than quantum, and this explanation involves the phenomena of *decoherence* that we'll get into shortly. The crucial observation is that an entangled but pure state in some higher-dimensional multi-qubit space can appear to be a mixed state when looked at from the point of view of a lower-dimensional subspace. Such mixed states that may appear when restricting the density matrix to a subspace by (partially) tracing out the other part are referred to as *improper mixtures*, and these are clearly essentially quantum because they derive directly from a pure (though entangled) state of the system.

Take a situation where we only koot at part of the system. It might be that we can only measure certain qubits and not others and without being aware of it. This is frequently the case because systems interact continuously with their environment. Studying the quantum behavior of a system, requires extraordinary precautions to make sure

A qubit named Botzilla. Once upon a time there were two qubits who had nothing to do with each other and therefore the two were in a separable state, say $|+\rangle_B \otimes |1\rangle_A$. We have a two-qubit system where qubit $\mathrm{Botzilla}$ is our object of study, while qubit $\mathrm{Abigail}$ is the girl out there who wants to get entangled (a form of quantum common-law marriage, so to speak) with our beloved $\mathrm{Botzilla}$. $\mathrm{Abigail}$, having studied equation (II.1.17), decided to bring in her charming friend CNOT to make it happen. If you lead us through the Gate, eternal gratitude will be yours! And this is what happened. Both of them, not so young, not-lovers really, went through the Gate anyway, and came out entangled indeed. As you may have anticipated, they did not live a long and happy life ever after. $\mathrm{Abigail}$ turned out to be a Botwoman and managed to one day disappear from the air, leaving $\mathrm{Botzilla}$ behind in a severely mixed-up state (basically making him the classical example of a quantum divorcee).

To understand the deplorable state $\mathrm{Botzilla}$ finds himself in, we have to perform what is called a *partial trace* in the quantum jargon. The point is that he can make only observations which concern himself, though, whether he wants or not, he is still entangled with Botwoman $\mathrm{Abigail}$. This means that he only has a small subset of observables to his disposal of the type $\mathcal{B} \otimes \mathbf{1} \in \mathcal{O}$. So calculating the expectation value of such an observable involves tracing over the $\mathrm{Abigail}$ qubit. This amounts to just establishing the fact that $\mathrm{Abigail}$ is still there' with unit probability, yet because the state is entangled, the effect of her 'being somewhere' is non-trivial. it leads to a result which can be described by saying that $\mathrm{Botzilla}$ is calculating the expectation value of just the operator $\mathcal{B}$ in his own system, but in a particular mixed state. So, he may ignore $\mathrm{Abigail}$, but then has to pay the price of being in a mixed state. Let us now 'trace out' $\mathrm{Abigail}$ and see what trace she left on $\mathrm{Botzilla}$'s state. This is achieved by summing over all the states associated with the subspaces we want to ignore, or better, about which we know nothing in particular. This means that we have to add up the diagonal entries with $\mathrm{Abigail}$ indices. We know already that $\mathrm{Botzilla}$ and $\mathrm{Abigail}$ ended up in the entangled state of equation (II.1.15), which we have to trace with respect to the second ($\mathrm{Abigail}$) qubit. This we do by making use of the fact that $\mathrm{tr}(\mathbf{1}\,|\psi\rangle\langle\phi|) = \langle\psi|\phi\rangle$. Using labels A and B to keep the qubits apart, and remembering that because we are using orthogonal basis states the calculation can be written like,

$$\begin{aligned}
\rho_B &= \mathrm{tr}_A\left(|\psi_{BA}\rangle\langle\psi_{BA}|\right)\\
&= \frac{1}{2}\mathrm{tr}_A\big[\,(|1\rangle_B|1\rangle_A + |-1\rangle_B|-1\rangle_A)\\
&\qquad\qquad (\langle -1|_A\langle -1|_B + \langle 1|_A\langle 1|_B)\,\big]\\
&= \frac{1}{2}\,(|1\rangle_B\langle 1|_B\langle 1|1\rangle_A + |-1\rangle_B\langle -1|_B\langle -1|-1\rangle_A)\\
&= \frac{1}{2}\,(|1\rangle_B\langle 1|_B + |-1\rangle_B\langle -1|_B)\,. \qquad \text{(II.1.26)}
\end{aligned}$$

This is the density matrix for $\mathrm{Botzilla}$ in a mixed state with probability $1/2$ to either be spinning up or spinning down. The corresponding entropy is also higher: In base-two $S = -\log(1/2) = 1$ bit, while for the original pure state $S = \log 1 = 0$. The whole operation of tracing out $\mathrm{Abigail}$ is non-unitary and irreversible, as we moved from two qubits to one. Indeed, exactly one bit of information got lost to the environment (it was taken along by $\mathrm{Abigail}$). In fact we could of course also calculate the entropy for the state $\mathrm{Abigail}$ finds herself in, then we have to trace over $\mathrm{Botzilla}$'s states. The situation is entirely symmetric, and her entropy will also be 1

bit. So there is some justice after all! This is what happens, the system $\mathrm{Botzilla} + \mathrm{Abigail}$ is in a pure state all along and the total entropy remains zero therefore, but looking at subsystems this is no longer true. What remains true is that if we divide the system up into two complementary parts, the entropy in each of them increases equally.

Generally it is the case that if we begin with a statistically pure separable state and perform a partial trace we will still have a pure state, but if we begin with an entangled state, and we perform a partial trace, we will get a mixed state as we just saw. In the former case the entropy remains zero, and in the latter case it increases. It is precisely in this sense that the Von Neumann entropy yields a useful measure of entanglement.

This observation is relevant for the understanding of real quantum systems, because most realistic quantum systems are strongly entangled with their environment. We don't know exactly how and with what, but it means that we tacitly trace out all kinds of things we are not aware of. What we know is that these systems behave quite classically in the end, and that in fact we should not be too surprised about that because they are in a strongly mixed state. □

that it does not engage in interactions that we have no control over. Such 'unknown knowns' might well wash away the quantum effects we were looking for. Quantum effects depend on the subtle phase relations that make quantum states in fact highly coherent. What to do if part of our system is out of sight? It boils down to a quantum, yet touching variant of the Romeo and Juliet story called 'A qubit named Botzilla.'

Event horizons revisited. The $\mathrm{Botzilla}$ tale we have just worked through may have reminded you of the black hole information paradox,which we addressed in the section on black holes in Chapter I.3 on page 139. We know that the Hawking-Bekenstein analysis leads to a macroscopic black hole entropy and temperature of the horizon. And we discussed that this is a property that can be assigned in a frame of reference where an event horizon is perceived. Our discussion of quantum entropy clearly allows for a microscopic mechanism, generating the entropy. We imagine the creation of a particle-antiparticle pair in a pure maximally entangled state, where one of the two particles falls through the horizon. This means that the Hilbert space factor corresponding to the lost particle gets traced out, which in turn tells us that the left-over particle finds itself in a mixed (maximal) entropy state. Very much like the $\mathrm{Botzilla}$ story. The entropy is the quantum entropy that arises because we are forced to take a partial trace. I have to admit that whether and how this perspective would fit into the 'quantum gravity' scenarios is still under serious debate.

Decoherence

Decoherence is the effect that a quantum system in a pure state loses its quantum coherence due to interaction with a complicated environment. It is one of the reasons why the world around us obeys the laws of classical physics.

Of course a quantum system may be in a pure state but if we do not take care it may quickly, through random interactions with the environment, end up in a mixed state. It is basically in a classical state where there are no quantum interference effects left and probabilities add, not quantum amplitudes. The quantum state 'decoheres'.

If we talk about qubit systems, then a way to think of these interactions is of course to think of gates that affect the

state and thereby cause decoherence. For example we may have a qubit in the $|+\rangle$ state, and have it interact with some phase gates, like a photon going through a random sequence of phase plates. The action of the phase-gate $P_z(\theta)$ corresponds to the unitary operator:

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \quad \rightarrow \quad \begin{pmatrix} 1 & 0 \\ 0 & e^{i\theta} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \alpha \\ e^{i\theta}\ \beta \end{pmatrix} .$$

Let us see what happens to the corresponding density matrix:

$$\rho_0 = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \rightarrow \rho(\theta) = P_z\ \rho_0\ P_z^* = \frac{1}{2}\begin{pmatrix} 1 & -i\theta \\ e^{i\theta} & 1 \end{pmatrix} .$$

Next we randomize the phases with some normal distribution as to represent 'the environment'. This means that we choose the density of dephasing agents to be Gaussian

$$f(\theta) \simeq e^{-\theta^2/\lambda} , \tag{II.1.27}$$

and then the effect of the random sequence of gates is obtained by averaging the above expression:

$$\sqrt{\frac{2i}{\lambda\pi}} \int f(\theta)\rho(\theta)d\theta = \frac{1}{2}\begin{pmatrix} 1 & e^{-\lambda/4} \\ e^{-\lambda/4} & 1 \end{pmatrix} \Rightarrow \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} .$$

What this calculation shows is that only the classical probabilities on the diagonal are left and the off-diagonal phase coherence of the quantum state ρ_0 has disappeared. By choosing λ large enough we wash out all quantum correlations and end up with a classical distribution over up and down states. This calculation merely illustrates a mechanism that leads to decoherence. Clearly, one would like to actually compute also the time-scales over which this decoherence takes place, this depends of course on the details of the environment or measurement apparatus.

Let us close this section by another toy model of decoherence. We start with a separable two-qubit state which we entangle using the CNOT gate as we did in (II.1.17). Then we use the $\mathrm{Botzilla} - -\mathrm{Abigail}$ mechanism by taking the partial trace with respect to Abigail as in (II.1.26) ending up with the mixed state for Botzilla. This basically turns the story into a decoherence phenomenon.

In other words, we imagine an interaction of a qubit B in a state $|\psi_B\rangle$ with the environment (a qubit A in some state $|\psi_A\rangle$) to generate an entangled two-qubit state $|\psi\rangle = |\psi_{BA}\rangle$ from a separable two-qubit state $|\psi\rangle = |\psi_B\rangle \otimes |\psi_A\rangle$. When viewed from the perspective of a single qubit, the resulting state after tracing out the A qubit, becomes incoherent. That is, suppose we look at (II.1.17) in the density matrix representation. Looking at the first qubit only, the state vector of the separable state is $|\psi_B\rangle = |+\rangle$,, a pure state in the density matrix representation given by equation (II.1.23),

$$|\psi_B\rangle\langle\psi_B| = |+\rangle\langle+| \Leftrightarrow \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} .$$

Under the action of CNOT this becomes the maximally entangled state on the right-hand side of equation (II.1.17). After partially tracing the density matrix as in (II.1.26) we end up with the B qubit in a mixed state given by (II.1.24),

$$\rho = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} .$$

Only the 'classical' probabilities on the diagonal are left and the off-diagonal phase coherence of the quantum state has disappeared due to entangling a degree of freedom in the environment.

Table II.1.1: **Key quantum principles concerning the Hilbert space of quantum states introduced in Chapter II.1.**

	Keyword	*Description*
(i)	**Hilbert space**	The complex vector space denoted by $\mathcal{H}$ of states of a quantum states. In this chapter we restrict ourselves to the finite dimensional case. We refer to the *Math Excursion* on complex vectors and matrices on page 632 of Volume III
(ii)	**State vector**	A *pure* quantum state is denoted by $\|\psi\rangle$ corresponding to a *ket* or column vector in $\mathcal{H}$.
(iii)	**Expansion of state in a basis**	Any state $\|\psi\rangle$ has a linear expansion in the basis $\{\|i\rangle\}$ given by $\|\psi\rangle = \sum_i \alpha_i \|i\rangle$, with normalization condition $\sum_i \|\alpha_i\|^2 = 1$.
(iv)	**Probabilistic interpretation**	A measurement on the above state $\|\Psi\rangle$ of the 'property' related to a basis $\{\|i\rangle\}$ gives outcome i with probability $\|\alpha_i\|^2$.
(v)	**Qubit state**	A qubit state is a two-dimensional complex vector: $\|\Psi\rangle = \alpha\|1\rangle + \beta\|-1\rangle$ with normalization $\|\alpha\|^2 + \|\beta\|^2 = 1$. A realization is a spin 1/2 degree of freedom where the vector is called a spinor.
(vi)	**Conjugate states**	The complex conjugate or dual state of $\|\psi\rangle$ is defined by the *bra* or row vector $\langle\psi\| = \sum_i \alpha_i^* \langle i\|$.
(vii)	**Bracket or inner product**	For two states $\|\Psi\rangle$ and $\|\Phi\rangle$ with coefficients α_i and β_i respectively, we define the inner product as the bracket $\langle\Phi\|\Psi\rangle = \sum_i \beta_i^* \alpha_i$. The orthonormal frame satisfies $\langle i\|j\rangle = \delta_{ij}$. $\langle\Phi\|\Psi\rangle$ is a complex number that satisfies $\langle\Phi\|\Psi\rangle = \langle\Psi\|\Phi\rangle^*$.
(viii)	**Multi-particle or qubit states**	If particle one has a state that is m-dimensional and that particle two is n-dimensional, than the two-particle system has a $(m \times n)$-dimensional state vector, which can be expanded as $\|\Psi^{(1,2)}\rangle = \sum_{i,j}^{M,N} \gamma_{ij} \|i^{(1)}\rangle \otimes \|j^{(2)}\rangle$. A two-qubit state vector is $2^2 = 4$-dimensional, written as: $\|\Psi\rangle = \alpha_1\|1,1\rangle + \alpha_2\|1,-1\rangle + \alpha_3\|-1,1\rangle + \alpha_4\|-1,-1\rangle$, with $\|i,j\rangle = \|i\rangle \otimes \|j\rangle = \|i\rangle\|j\rangle$.
(ix)	**Entangled and separable states**	A n-particle state is *separable* if it the state can be factorized in an n-fold product: $\Psi^{(1,2,\ldots,n)}\rangle = \|\psi^{(1)}\rangle\|\psi^{(2)}\rangle \cdots \|\psi^{(n)}\rangle$. A state is *entangled* if it is not separable.
(x)	**Mixed states**	A mixed state is a properly normalized (statistical) mixture of some set $\{\|\psi_a\rangle\}$ pure states: $\|\Psi\rangle = \sum_a p_a \|\psi_a\rangle$, with probability p_a that the system is in the pure state $\|\psi_a\rangle$.
(xi)	**Density matrix/operator**	The density operator for a mixed state $\|\Psi\rangle$ is defined as $\rho = \sum_i p_a \|\psi_a\rangle\langle\psi_a\|$ For a pure state there is only one term $p = 1$.
(xi)	**Quantum entropy**	The quantum entropy of a mixed state is given by: $S(\rho) = -\mathrm{tr}\,\rho \log \rho = -\sum_a p_a \log p_a$. For a pure state the entropy is zero.

Chapter II.2

Observables, measurements and uncertainty

> It is wrong to think of that past [ascribed to a quantum phenomenon] as 'already existing' in all detail. The past is theory. The past has no existence except as it is recorded in the present. By deciding what questions our quantum registering equipment shall put in the present we have an undeniable choice in what we have the right to say about the past.
>
> *John Archibald Wheeler,*
> *Some Strangeness in Proportion* (1980)

In the previous chapter we focussed exclusively on states, in particular the space of pure quantum states, the Hilbert space $\mathcal{H}$. In this chapter we consider the physical variables or quantum observables. These are represented by linear operators or matrices which act on the Hilbert space. The fact that physical variables are no longer represented by ordinary numbers or functions like in classical physics, but by matrices or differential operators makes quantum theory fundamentally different. It leads to deep reflections on the logical structure of the theory, on the nature of measurements, and on the fundamental aspects of uncertainty so concisely expressed by the Heisenberg uncertainty relations. And that is what this chapter is about. It should make you feel at home in Hilbert space.
We have summarized and specified the basic ingredients of the mathematical framework and the jargon that comes along with the notion of observables, which forms the subject of this chapter, in the table at the end of the chapter on page 321.

Quantum observables are operators

Physical variables or observables in quantum theory are represented by hermitian operators. In this section we explore what this means in general and work out most of the details for the case of qubits. Operators have a spectrum of eigenvalues that correspond to possible measurement outcomes. To these eigenvalues correspond orthogonal eigenstates (or subspaces), which can be used to define a suitable frame for the Hilbert space. The aim of this section is to exhibit the algebraic structure of the theory, with the observables, projection operators and raising and lowering operators which play essential roles in describing the generic properties of quantum systems.

The algebra of observables. For a quantum system we have a set of dynamical variables called *observables*, $\mathcal{O} = \{\mathrm{A}, \mathrm{B}, \ldots\}$. In most cases corresponding to the classical variables, but there may be additional variables such as the aforementioned spin, which have no classical analogue. Whereas in classical physics the language of states and dynamical variables is smoothly connected, basically because the states are labelled by the (real) values of the

dynamical values. This however is no longer true in the quantum world. In quantum theory we make a clear distinction between the Hilbert space $\mathcal{H}$ of states and the set of observables $\mathcal{O}$. Let us start with some general properties and definitions.

1. Operators on Hilbert space. The quantum observables are represented by linear *operators*, that act on Hilbert space.[1] In other words we have that $\mathcal{O} : \mathcal{H} \to \mathcal{H}$, and we write:

$$|\psi'\rangle = A\,|\psi\rangle\,,$$

with $|\psi'\rangle, |\psi\rangle \in \mathcal{H}$ and $A \in \mathcal{O}$.
You should typically think of *matrices* in case the Hilbert space is finite dimensional.[2] In the infinite-dimensional case, we should think of continuous systems like a particle, where the states are described by wavefunctions $\psi(x,t)$, and the operators are typically represented by a differential operator, like the momentum and energy operators:

$$P = -i\hbar\frac{d}{dx} \quad \text{and} \quad H = i\hbar\frac{d}{dt}\,,$$

as we mentioned in the previous chapter.
The fact that observables are operators that 'act on states' implies that they may well change the physical state, and strongly suggests the possibility that the act of measurement of such an observable will affect the state of the system.

2. Linearity. Linearity implies that for any two states and any observable A we have that,

$$A\,(|\psi_1\rangle + |\psi_2\rangle) = A\,|\psi_1\rangle + A\,|\psi_2\rangle\,.$$

3. Hermitian adjoint. On the algebra $\mathcal{O}$ we can define a hermitian adjoint, or 'dagger' operation, denoted as $\dagger$, where $A \to A^\dagger$. The definition is as follows

$$\langle\phi|A^\dagger|\psi\rangle = \langle\phi|A|\psi\rangle^* \quad \text{for all} \quad |\phi\rangle, |\psi\rangle \in \mathcal{H}\,. \qquad \text{(II.2.1)}$$

Sandwiching the adjoint operator $A^\dagger$ between any pair of states yields a number, which is the complex conjugate of the number resulting from sandwiching A. From the definition it follows that (i) the adjoint of a product satisfies $(AB)^\dagger = B^\dagger A^\dagger$, and (ii) the dagger squares to unity: $(A^\dagger)^\dagger = A$, and is therefore referred to as an *involutive automorphism* of the algebra of observables. For matrices this implies that the hermitian adjoint of A is defined as $A^\dagger = (A^{tr})^*$, or in words: it is the complex conjugate of the transpose of A.

4. Hermitian or self-adjoint operators. We require that the eigenvalues of an observable are real numbers, as they correspond to possible outcomes of measurements, and that translates into conditions on the particular type of matrices that can represent physical observables. As a matter of fact the reality condition on the eigenvalues of operators requires that the quantum observables have to correspond to *hermitian* also known as *self-adjoint* operators or matrices. This means that observables satisfy the condition $A = A^\dagger$. A general hermitian matrix is a matrix M with complex entries that can be written as $M = S + iA$, where S is real and symmetric, and A is real and antisymmetric. For the case of a two-dimensional Hilbert space, like in the case of a single qubit or a basic quantum spin, all observables can be expressed as a linear combination of the unit matrix and the three Pauli or spin matrices of equation (II.2.2).

5. Norm and boundedness. We like to talk about bounded operators A, meaning that if they work on vectors in Hilbert space they do decent things. So what sets the norm $||A||$ for an operator? Here is a reasonable way to do this: (i) you let A work on all states in $\mathcal{H}$, (ii) calculate the norms of all the resulting vectors, and (iii) look at the 'largest value' or 'infimum' that occurs, which is denoted

[1] In this book we adopt the convention to represent quantum observables with uppercase letters while for their values we use lowercase. The set $\{a\}$ of allowed values is called the *sample space* of the observable A.

[2] We refer to the *Math Excursion* on page 614 of Volume III for an introduction to real matrices and vectors, which was extended to the complex case in the *Math Excursion* on page 632.

by $\inf$. So, the definition of the norm of the operator A is then:

$$\|A\|^2 = \inf\{\langle\psi|A^\dagger A|\psi\rangle : \forall|\psi \in \mathcal{H}\}.$$

A *bounded operator* has by definition a finite norm: $\|A\| < \infty$. If you think of A as a matrix, this statement boils down to saying that the eigenvalues of the matrix should be finite.

6. Algebraic structure. The observables form an algebra (we want to add and multiply observables). This is easy to understand for matrices as we will show in the *Math Excursions* just mentioned. The restrictions (of boundedness and self-adjointness) are much harder to implement if one passes to the infinite-dimensional cases corresponding to physical systems like particles and fields which have continuous variables. To properly address these problems one needs some quite sophisticated mathematics involving concepts like *Banach spaces* and C^* *('C-star') algebras*. This allows for a mathematically rigorous and consistent formulation of quantum theory. Such axiomatic approaches, however, are far beyond the scope of this book, though one may of course argue that they are quantessential because they address foundational questions. We will follow an operational, less rigorous approach, and comfortingly, it turns out that the typical notation we have introduced doesn't change much after going rigorous. We will treat the expressions using simple rules, glossing over the fact that we manipulate symbols which deep down may refer to rather sophisticated notions.

The qubit observables. The Hilbert space for a qubit is two-dimensional, and therefore the observables can be represented by 2×2 hermitian matrices. A typical set of observables would be the set of so-called Pauli matrices $\{X, Y, Z\}$ with:

$$X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \; Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \; Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad \text{(II.2.2)}$$

Any one qubit observable can be expressed as a linear combination of the three Pauli matrices and the unit matrix.[3]

In our discussion of classical bit mechanics we already argued that the X matrix, as operator or gate, acts like a momentum or displacement operator on the z-space of the bit, because it acts like the NOT-gate interchanging the two bit states $|1\rangle \leftrightarrow |-1\rangle$. It shows nicely how classical physics (discrete mechanics), and now quantum theory meet in this picture, with a correspondence between dynamical maps, logical (digital) gates, and quantum observables: they are all operators acting on a state.

q-gates. Clearly the three Pauli spin matrices above are one-qubit gates. In classical computation the X-gate corresponds to the NOT-gate, and is the only acceptable one-bit gate. The others are not, because the Z-gate introduces a relative minus sign (which is a phase), and the Y-gate introduces complex components, which are both not admissible for classical bits. This is a first hint that quantum bits offer far more possibilities, so let us get back to the qubit observables.

Sample spaces and preferred states

To each observable A corresponds a set $\mathcal{S}_a = \{a_i\}$ of values it can take. In other words, it is the set of possible outcomes of a measurement of the observable A, which is also called the *spectrum* or *sample space* of A. If we apply the observable A to a state $|\psi\rangle$ and we get a number a_i multiplying that same state, we say that the system is in a state where A takes the value a_i. A state with this property is denoted as $|\psi\rangle = |a_i\rangle$, and is called a *preferred* or *eigenstate* (or eigenvector) of A with *eigen-*

[3] The real spin polarization operator has units and equals $S_z \equiv \frac{1}{2}\hbar Z$, involving an essential factor one half. Throughout the book we discuss spin one-half directly in terms of the Pauli matrices $\{X, Y, Z\}$, which in most textbooks are denoted as $(\sigma_x, \sigma_y, \sigma_z)$.

value a_i. These statements are summarized by the following equation,

$$A|a_i\rangle = a_i|a_i\rangle\,. \qquad \text{(II.2.3)}$$

Is the eigenvector defined this way unique? No, it is not, we can multiply by any overall constant and it is still an eigenvector. We take care of that by choosing the eigenvector to have unit length, but then there is still an overall phase factor $(e^{i\phi})$ possible. This factor doesn't have any observable consequences.

Qubit eigenstates. Recall that for the classical dynamical bit we introduced a position $z = \pm 1$ and a momentum $p = \pm 1$. In the quantum realm these observables should somehow correspond to certain operators. Let us thereto consider the 2×2 matrices Z and X (related to p) which can act on the states in $\mathcal{H}$.

The basis vectors corresponding to the classical states are indeed eigenvectors of the position operator Z:

$$Z\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix} = 1\begin{pmatrix}1\\0\end{pmatrix},$$

$$Z\begin{pmatrix}0\\1\end{pmatrix} = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\begin{pmatrix}0\\1\end{pmatrix} = -1\begin{pmatrix}0\\1\end{pmatrix},$$

and the eigenvalues $z_\pm = \pm 1$ are the corresponding z values. We conclude that the sample space or spectrum of the observable Z is $\mathcal{S}_z = \{\pm 1\}$.

The operator X does also exactly what you would expect of the 'momentum' operator; it implements the $p = 1$ transition $|\pm 1\rangle \Leftrightarrow |\mp 1>$ as one may verify explicitly:

$$X\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}0\\1\end{pmatrix},\ \text{etc.}$$

We also learn that the operator X^2 equals the unit matrix. In fact we have that $X^2 = Y^2 = Z^2 = \mathbf{1}$, which by definition leaves all states invariant and it therefore implements the trivial $p = 0$ transition. This is as far as the 'relation' between classical and quantum formalism can be traced.

The quantum formalism allows for more because we have the *linear superposition principle* as well as the *complexification* of the state vectors. We have seen that the states $|\pm 1\rangle$ correspond to the eigenvectors of the 'position' operator Z, but in the quantum formalism we can also ask for the eigenvectors of other observables, for example X. One easily verifies that these correspond to the state vectors $|\pm\rangle = (|+1\rangle \pm |-1\rangle)/\sqrt{2}$, with again eigenvalues $x_\pm = \pm 1$ as follows:

$$X\begin{pmatrix}1\\ \pm 1\end{pmatrix} = \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}\begin{pmatrix}1\\ \pm 1\end{pmatrix} = \pm 1\begin{pmatrix}1\\ \pm 1\end{pmatrix}. \qquad \text{(II.2.4)}$$

The eigenvectors $|\pm\rangle$ are real linear superpositions of the basis states $|\pm 1\rangle$, and we have marked them on the circle of real states in Figure II.1.7.

Is this all? Are we done? The answer is, no! We have indeed identified the eigenstates of momentum, which actually do not have a classical equivalent. This shows the quantessential possibility that the linear superposition principle introduces. However, we have so far only explored real states and real matrices, and it is here that the quantum formalism summons us to proceed. There are other independent choices: the one conventionally chosen is the (complex) matrix Y:

$$Y = \begin{pmatrix}0 & -i\\ i & 0\end{pmatrix}.$$

One may verify that $Y = Y^\dagger$, and we see that acting on the basis states it indeed introduces complex coefficients as

$$Y|\pm 1\rangle = \pm\, i\,|\mp 1\rangle\,.$$

So loosely speaking we could say that Y introduces a complex part to the standard classical momentum variable $P \simeq$

X. We should expect its eigenstates to be complex as well:

$$Y\begin{pmatrix}1\\ \pm i\end{pmatrix}=\begin{pmatrix}0 & -i\\ i & 0\end{pmatrix}\begin{pmatrix}1\\ \pm i\end{pmatrix}=\pm 1\begin{pmatrix}1\\ \pm i\end{pmatrix},$$

and the eigenvalues are again $y_\pm = \pm 1$.

The fact that all eigenvalues square to one is not surprising if one realizes that the matrices themselves square to the unit matrix: $Z^2 = P^2 = Y^2 = \mathbf{1}$. All quantum observables in this problem can be written as linear combinations of the independent hermitian matrices X, Y, Z and the unit matrix $\mathbf{1}$. These basic observables have identical sample spaces $\mathcal{S}_x = \mathcal{S}_y = \mathcal{S}_z = \{+1,-1\}$. Furthermore, as we just showed, they have no eigenvectors in common. It signals the important fact that these three observables are incompatible with each other, a notion we will return to later on. It raises the question of what that means in terms of measuring these observables in such a non eigenstate.

Expectation values. We may now also define the notion of the expectation value of an observable A in a quantum state $|\psi\rangle$ as:

$$a =< A >\equiv \langle\psi|A|\psi\rangle\,, \qquad \text{(II.2.5)}$$

which is just a number indeed. The expectation value a is therefore a weighted average of the eigenvalues of A, which depends on which state $|\psi\rangle$ one chooses. This is consistent with the remark we made earlier that the square of the coefficients are probabilities. It means that we 'sandwich' the operator between a row and column vector, for example:

$$\langle +1|\,Z\,|+1\rangle = (1\,0)\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\begin{pmatrix}1\\ 0\end{pmatrix}=(1\,0)\begin{pmatrix}1\\ 0\end{pmatrix}=1\,,$$

and similarly:

$$\langle +1|\,X\,|+1\rangle = \langle +1|-1\rangle = 0\,.$$

An expectation value can be calculated for any observable in any state and corresponds to some average of measurement outcomes.

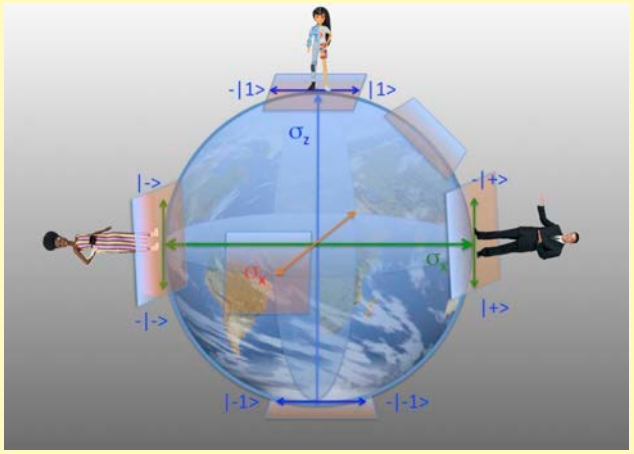
A Qubit is like a Barbie on a globe

We return to the qubit state space and point out an alternative way to parametrize qubit space by directly relating it to the eigenstates of an operator/observable. This amounts to yet another geometrical representation of the state space of a qubit or quantum spin, and that will be useful in a variety of contexts. We start by choosing a point on the unit two-sphere in X, Y, Z space, as depicted in the figure (a). The point represents a unit vector $\hat{n}$, but we want to use it to label a qubit state, which as we saw is a point on a unit three-sphere so we have to do a little more. First we construct a unit sigma matrix $\hat{n}\cdot\vec{\sigma}$, with $\sigma = \{X, Y, Z\}$, which to each point on the sphere associates a particular hermitian ($2\times$ 2) matrix or observable. This observable is proportional to the spin operator along that axis. The qubit state that we link to the point is the eigenvector $|\chi_1\rangle$ of that observable with the highest eigenvalue ($\lambda_1 = 1$). However the eigenvector is not unique: it is multiplied by a phase factor with some angle ϕ between zero and 2π, so to completely fix the state we have to specify the pair $\{\hat{n}, \phi\}$.

The mathematically alert reader may have experienced a feeling of déjà vu since I am basically repeating the story I told in Chapter I.1 about the Hopf or monopole bundle, where the three-sphere was interpreted as a phase or circle bundle over the two-sphere. So the three-sphere is a physically relevant object, we have seen it appear as the bundle associated with the fundamental Dirac monopole in Chapter I.1, as the manifold of the group $SU(2)$ in the *Math Excursion* in Volume III on Groups, and here as the state space of a qubit.

The natural way to represent also the angle ϕ

in the picture is to draw the tangent plane to the sphere at the point chosen, and define ϕ as the polar angle in that tangent plane as we did in the *Math Excursion* on Complex numbers on page 607 of Volume III.



(a): *Choosing a state of a qubit is like setting a Barbie on a globe.*. Choosing a different frame is like choosing the North and South Poles along a different axis.

In the figure we have depicted some of the states we discussed before, on the z-axis we have two points with the operators $\pm Z$ with eigenvectors $|\pm 1\rangle$. So the states are now represented as unit vectors in the tangent plane at the point $\hat{n}$ with phase angle ϕ. So in the plane at the North Pole we find the states $\pm|1\rangle$ at angles $\phi = 0, \pi$.

What we have learned is that we can represent a point on the three-sphere by choosing a point on the two-sphere and an additional phase in that point. This way of choosing coordinates on the three-sphere is indeed completely equivalent to fixing a Barbie on the earth surface by saying *where* (s)he stands, *and* in *what direction* (s)he is looking. In a more sophisticated wording one says one picks a point on the sphere and a frame in the tangent plane to the sphere at that point as is illustrated in the figure. So now you don't any longer have to say that you cannot imagine how to choose a point on a three-sphere, even a kid can do it! Buy him a Barbie of some sort and a globe and ask him to stick the Barbie on the Globe.

Note that the present picture (a) is essentially different from Figure II.2.2 in that corresponding states are located in different places. For example the North Pole represents the states $\alpha\ |1\rangle = \exp(i\phi)|1\rangle$, where ϕ is the angle of the arrow in the tangent plane.

This set contains in particular the real states $|1\rangle$ for $\phi = 0$ and $-|1\rangle$ for $\phi = \pi$, whereas the states $\pm|-1\rangle$ are located on the South Pole. In Figure II.2.2 the states $|\pm 1\rangle$ are perpendicular, in Figure II.2(a) they are antipodal. Changing the qubit state corresponds to moving around on the three-sphere and that is nothing but walking over the globe and looking in various directions. What is all this good for? This alternative view of the space of states of a qubit or quantum spin has yielded some interesting physical insights to be addressed in Chapter II.3 about probing the state space and measuring the Berry phase, which is exactly like having the Barbie in the figure walking around on the globe. □

Spin or qubit Hamiltonians

A crucial observable in physics is the energy or the Hamiltonian operator denoted by H. The eigenvalues E_n of the Hamiltonian correspond to the allowed energy levels of the system. The possible energy eigenstates $|\psi_n\rangle$ are called *stationary state*s, because they have a trivial time

dependence that resides in the overall phase factor. Linear combinations of different energy eigenstates would therefore have a non-trivial time dependence. Of particular interest is the lowest energy state or ground state of the system. We consider two examples for the Hamiltonian of a spin or qubit and show their properties. Our first choice corresponds to putting the spin in a magnetic field in the z-direction, the Hamiltonian would be proportional to Z:

$$H_1 = bZ\,,$$

and its eigenstates are $|\pm 1\rangle$ and have eigenvalues $\lambda_\pm = \pm b$. Another sensible choice for the Hamiltonian would be what is called the total spin operator which is quadratic in the spins:

$$H_2 = b(X^2 + Y^2 + Z^2) = b(\mathbf{1}+\mathbf{1}+\mathbf{1}) = 3b\mathbf{1}\,,$$

Indeed a bit trivial perhaps, because it is just 3 times the unit matrix. Of course, if we act with this Hamiltonian on *any* state, it will return that state with eigenvalue $3b$, i.e. $H_2|\psi\rangle = 3b|\psi\rangle$. In this case you could say that the Hamiltonian is trivial, because all states have the same eigenvalue, they are what we call *degenerate*. Degeneracies are a common feature and usually imply that there is some (hidden) symmetry in the system one considers.

Frames and observables

The *eigenstates* $|a_k\rangle$ of a linear operator A are defined by the equation $A|a_k\rangle = a_k|a_k\rangle$. If A is a $N \times N$ hermitian (matrix) operator, there are N independent (N-dimensional) eigenvectors and the *eigenvalues* a_k are real and generically different. As we will see these eigenvalues are the possible outcomes of a measurement of that observable. Generally the eigenstates can be chosen orthonormal, so that

$$\langle a_j|a_k\rangle = \delta_{jk} \text{ where } \delta_{ij} = 1 \text{ if } i = j\,, \text{ and } \delta_{ij} = 0 \text{ if } i \neq j\,. \tag{II.2.6}$$

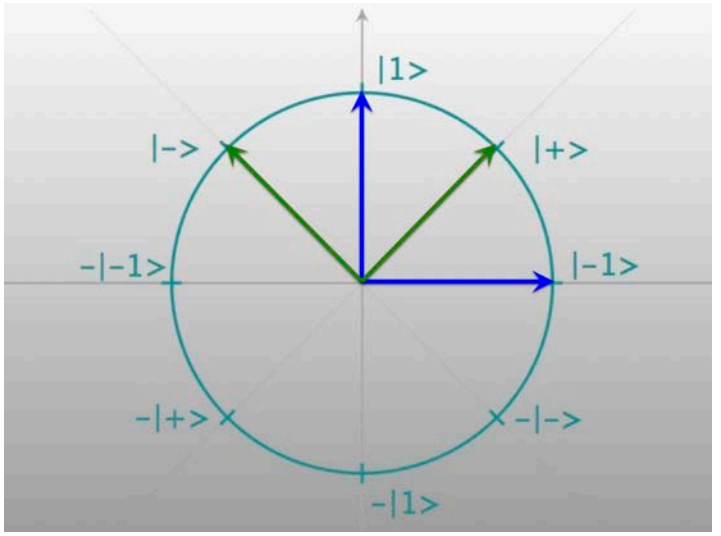


Figure II.2.1: *Two frames.* Two different frames spanning the same two-dimensional space of real qubit states. The blue one is the Z frame $\{|-1\rangle, |1\rangle\}$ and the green one is the X frame $\{|+\rangle, |-\rangle\}$. The frames are related by a rotation over an angle $\theta = 45^o$.

This means that the set $\{|a_i\rangle\}$ forms an *orthonormal basis* or *orthonormal frame* for the state space – the Hilbert space – of the system.

Qubit frames. Let us briefly illustrate this: the eigenstates for $A = Z$ are the column vectors

$$|1\rangle \Leftrightarrow \begin{pmatrix}1\\0\end{pmatrix}, \qquad |-1\rangle \Leftrightarrow \begin{pmatrix}0\\1\end{pmatrix},$$

which have eigenvalues plus and minus one respectively. The eigenstates $|\pm 1\rangle$ of Z form an orthonormal basis for the space of qubit states.

If we choose instead $A = X$, then the normalized eigenstates correspond to

$$|\pm\rangle \equiv \frac{1}{\sqrt{2}}(|1\rangle \pm |-1\rangle) \Leftrightarrow \frac{1}{\sqrt{2}}\begin{pmatrix}1\\\pm 1\end{pmatrix}$$

and these have eigenvalues ± 1 also. Clearly, the states $|\pm\rangle$ form an alternative basis for the qubit states. In Fig-

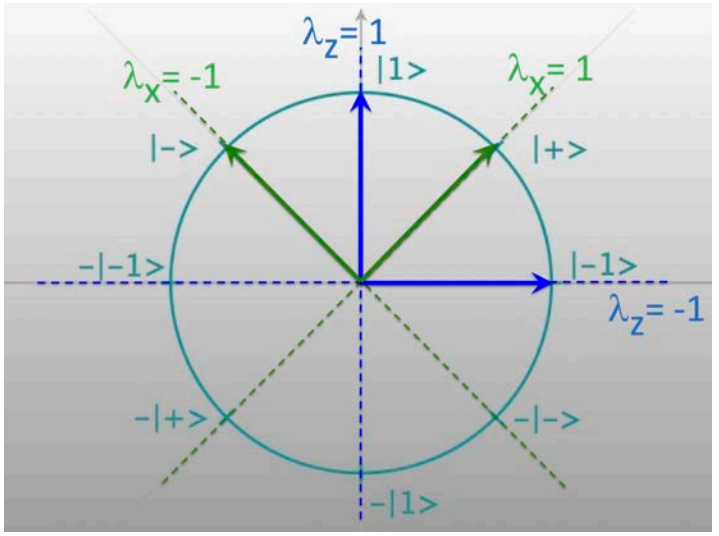


Figure II.2.2: *Frames and eigenvalues.* The frames corresponding to eigenstates of Z (blue) and X (green) respectively. The axes are labeled by the corresponding eigenvalues. The circle represents the normalized qubit states with real α and β.

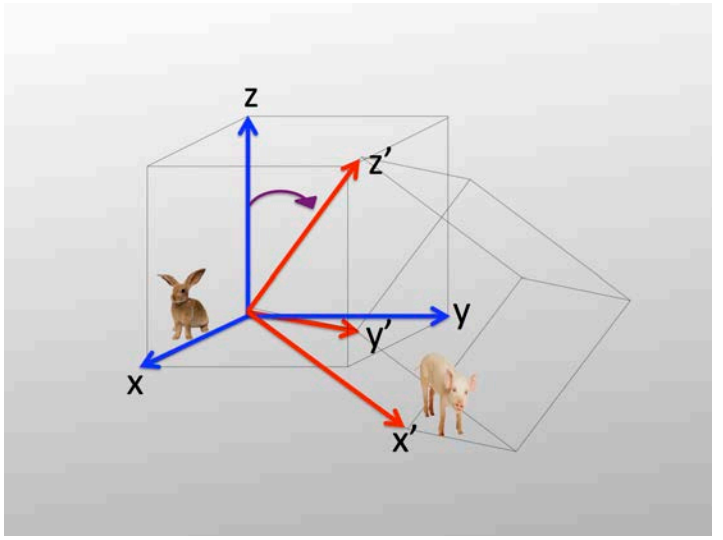


Figure II.2.3: *Frame rotations.* Two frames spanning the space $\mathbb{R}^3$. The 'rabbit' and the 'pig' frames can be rotated into each other. For example first rotate the z'-axis to the z-axis, then the $x-, x'-, y-$ and y'-axes all lie in the $x-y$ plane. So there they can then be rotated into each other by a rotation around the z-axis. This also holds for frames in higher dimensions because rotations preserve the origin, the length of vectors and also the angles between them. This in fact defines what a rotation is.

ure II.2.2 we have depicted the two frames where the unit circle describes all the states with real coefficients α and β. This picture will return in many guises when we discuss measurements in quantum mechanics. A priori there is no preference for any particular basis, the best choice depends on the questions you want to answer. Clearly if we are going to measure some physical quantity, the eigenstates of the corresponding operator will play an important role.

What the examples just given also show is that the Z and X operators have no eigenvectors in common. That is necessarily the case because the operators do not commute, and they are called *incompatible observables*. We return to this notion in a forthcoming section.

Frame choices. When writing down an explicit expression for a qubit, or in fact for any quantum system, we first have to choose a *basis* $\{|i\rangle\}$ in which the state can be expanded. This basis is a matter of choice. In Figure II.2.1 we have for example depicted the standard blue frame, but also a different green frame consisting of the states $|+\rangle$ and $|-\rangle$. In Figure II.2.3 we have depicted two frames for a three-dimensional vector space. What is quite evident from the figures is that different frames can be transformed into each other by a simple rotation. That is so because rotations by definition not only keep the length of vectors but also the angles between them the same.[4]

[4] A rotation in fact preserves the *orientation* of a frame. If we interchange the x- and y-axes in Figure II.2.3, then we also have an orthonormal frame, but it cannot be obtained by rotating the old frame, exactly because its orientation is opposite. The frames in the figure are *right-handed* meaning to say if rotate from x to y the right-handed rotation by the 'like'-rule would point in the positive z-direction.

Unitary transformations

A rotation of a vector or a frame is an operation or a transformation on such a vector or frame. You may in this respect think of a frame as a solid cube, under rotations its shape is conserved, it stays congruent. In a $N-$dimensional space such rotations can be represented by a $N \times N$ matrix that act on a vector.

An important property of rotations is that they satisfy the *group* property, namely that the result of two successive rotations is again a rotation. This is obvious in the two-dimensional case because you just add the angles. In three dimensions a simple way to see it is to look at the 'rabbit' unit vector in the $z-$direction in Figure II.2.3. If we would trace out the arrow head under all possible rotations, we should get the unit sphere. Any rotation of a vector around an *orthogonal* axis would move the arrowhead along a *big* circle over the sphere, big because it is a circle of maximal size on a given sphere. It is also true that the shortest distance between two points on the sphere is exactly the unique segment of the unique big circle on which both points lie. So, if we make first a rotation of the vector around some axis $\hat{n}_1$, the vector moves from the first point A over a segment of some big circle to a second point B. Next we move the resulting vector over a given angle around a second axis $\hat{n}_2$, then the vector ends up at a third point C on the sphere. The combined rotation is then just the rotation that moves the vector from A directly to C over the big circle connecting them.

This is all simple to imagine, and therefore let us now translate these simple geometric intuitions into a symbolic language. We start with rotating ket vectors with rotation matrices U_i:

$$\begin{aligned} |\psi_2\rangle &= U_1|\psi_1\rangle \\ |\psi_3\rangle &= U_2|\psi_2\rangle = U_2U_1|\psi_1\rangle = U_3|\psi_1\rangle \\ &\Rightarrow U_3 = U_2U_1\,. \end{aligned} \tag{II.2.7}$$

This is true for arbitrary vectors and also for arbitrary rotations. Under a frame rotation U, the conjugate bra vector will rotate like:

$$\langle\psi_2| = \langle\psi_1|\,U_1^\dagger\,,$$

with the conjugated rotation matrix $U^\dagger$, that can be obtained from U by interchanging rows and columns (which is called taking its *transpose* U^{tr}) and also taking its complex conjugate (meaning conjugating all its matrix elements i.e. its entries, so, $U^\dagger = (U^{tr})^*$. We require the length and inner product of vectors to be preserved under rotations, so if we simultaneously rotate arbitrary vectors $|\psi\rangle$ and $|\phi\rangle$ by U, then we have to impose:

$$\langle\phi_2|\psi_2\rangle = \langle\phi_1|U^\dagger\,U|\psi_1\rangle = \langle\phi_1|\psi_1\rangle\,.$$

From the last equality we conclude that rotations apparently correspond to a *unitary transformation*, satisfying the unitarity condition:[5]

$$U^\dagger U = \mathbf{1}\,.$$

The rotations in N complex dimensions form a mathematical structure called a *group*, basically because they satisfy the group property, equation (II.2.7). This group is called the *unitary group* denoted by $U(N)$. More precisely it is the *special unitary group* $SU(N)$ because the rotations preserve the *orientation* of the frame (this is the cyclic order X, Y, Z, where by definition $\hat{x} \times \hat{y} = \hat{z}$). We refer to the *Math Excursion* A for further details.

Photon gates and wave plates

One can think of these unitary operations as a transformation on the qubit state vector. And changing the state

[5] Note that if we rotate in real space the matrices become real and there is no complex conjugation, therefore real rotations are orthogonal matrices O satisfying the condition that $O^{tr}O = \mathbf{1}$ these matrices also form a closed group under multiplication, denoted as the orthogonal group $O(N)$. Indeed where quantum physicists are married to unitary groups, classical physicists are with the orthogonal ones. It is the difference between being complex and being real.

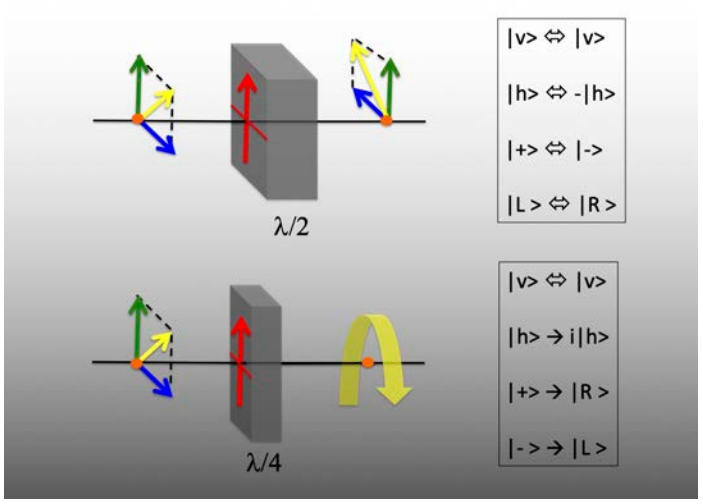


Figure II.2.4: *Wave plates.* The wave plates with optical thickness of $\lambda/2$ and $\lambda/4$ can be used to change the polarization state of a photon. They are unitary one-qubit phase gates, and are the physical realizations of the transformations U described in the text, on the states defined in Figure II.1.11. The transformations can be inverted meaning that we reverse the direction in the picture, so, if going to the right corresponds to some U, then going to the left corresponds to $U^\dagger$.

vector really amounts to processing quantum information as the *in-state* gets transformed into some *out-state*. Such manipulations can be performed on real photons relatively simply by what are called *wave plates*. These have two parameters: a principal axis and a given optical thickness as is depicted in Figure II.2.4. We have shown the effect on the polarization state of a photon when it passes through a phase plate with its principal axis along the z-axis in the figure. The plate acts like what is called a *phase-gate* $P(\theta)$; it leaves the polarization along the principal axis unchanged, and rotates the orthogonal component by a phase corresponding with the optical thickness of the plate. So in the case at hand the action is given by,

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \quad \rightarrow \quad \begin{pmatrix} 1 & 0 \\ 0 & e^{i\theta} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \alpha \\ e^{i\theta}\,\beta \end{pmatrix} .$$

Indeed the $\lambda/2$ plate rotates the lower component over an angle $\theta = \pi$ in the complex plane leading to the phase -1, while the $\lambda/4$ plate rotates by an angle $\theta = \pi/2$ giving an imaginary factor i.

Incompatible observables

The fact that observables are represented by operators reflects the quantessential property that measurements may alter the state, and therefore that the outcomes of different measurements may depend on the order in which the measurements are performed. This latter property expresses the fact that the operators that represent observables in quantum mechanics do not necessarily *commute*, by which we mean that for the product of two observables A and B one may have that $AB \neq BA$ and we say that such observables are *incompatible*. It is pretty weird to be told that $\mathrm{momentum}$ times $\mathrm{position}$ would not be equal to $\mathrm{position}$ times $\mathrm{momentum}$, but that is the way it really is if you think of them as operators instead of numbers. This is common in the quantum world because matrices generically do not commute. For the simple set of qubit observables given in equation (II.2.2), you can verify that they do not commute with another indeed: for example $ZX - XZ = 2iY$.

To illustrate this *non-commutativity* we have in Figure II.2.5 depicted a sequence of two 90^o rotations in opposite order: on the left we rotate the book first around the z-axis and then around the x-axis, and on the right we do it in the opposite order. At the bottom one sees that the resulting orientations of the book clearly differ, meaning that for the operations on the state of the book b one has that $R_z R_x \neq R_x R_z$. For the case of a particle it turns out that the position and momentum observables X and P do not commute: one finds that $XP - PX = i\hbar$. This non-commutativity of observables has dramatic consequences and lies at the root of many of the at first sight *inconvenient truths* that quantum theory revealed about the basic workings of nature.

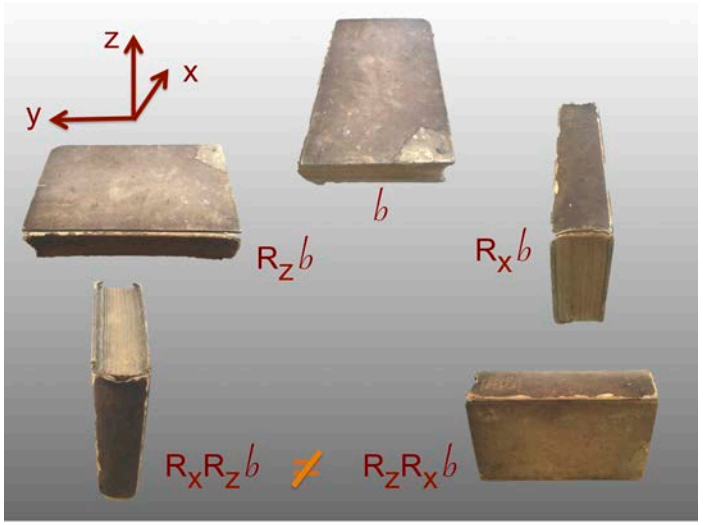


Figure II.2.5: *Non-commuting rotations.* We illustrate non-communativity of the 90^0 clockwise rotations R_z and R_x around the z- and x-axes respectively. The order in which they are applied (to the book) does matter and clearly leads to a different final state.

The labelling of quantum states. Consider a $\mathrm{N} \times \mathrm{N}$ matrix observable, in the generic case it will have N different real eigenvalues, with orthogonal eigenvectors. In general, it may happen that two or more eigenvalues coincide, in which case there will be more than a single (independent) eigenvector corresponding to a given eigenvalue. We say that the spectrum of the observable A is *degenerate*. In that case the eigenvalue $\mathrm{a_i}$ labels not just a particular state but rather some subspace $\mathcal{V}_{\mathrm{i}}^{\mathrm{a}}$ of the Hilbert space. In fact states can be simultaneous eigenstates of other observables. The previously mentioned state may also be an eigenvector with value $\mathrm{b_j}$ for the observable B, and we may label that state by the element of the combined sample space and write $|\Psi\rangle = |\mathrm{a_i}, \mathrm{b_j}, \ldots\rangle$.

In general there will be many different sets consisting of a maximal number of independent, but compatible observables and these can be used to label a particular set of basis states (a frame) of the system. Observables A and B for which a joint set of eigenstates can be chosen, necessarily commute and are therefore by definition *compatible*. What makes quantum theory so special is that this is often not the case, so that we continuously have to deal with observables A and B that are *incompatible*. For such incompatible observables Heisenberg's uncertainty relations impose quantessential restrictions, to which we will turn shorty.

Quantum setting. We conclude that there are four aspects in which the quantum setting significantly differs from the classical one:
(i) the set of admissible values for a dynamical variable may differ, in particular it may be a discrete set in which case the values would be quantized whereas in the classical case the values would be continuous;
(ii) a quantum variable may not have a classical analogue at all, such as a particle having an intrinsic rotational degree of freedom called 'spin', and most importantly;
(iii) in a given state of a quantum system generally *incompatible observables cannot be simultaneously assigned a definite value*. The non-zero spread in observed values in a given state is then governed by Heisenberg's uncertainty principle to be discussed later;
(iv) certain classical dynamical variables which involve products of incompatible variables will not have an unambiguous or unique quantum analog. There may be ordering ambiguities.

At first it seems inconceivable that such a vile theory has become one of the crown jewels of a rigorous science like Physics! It is remarkable that a theory can host this very anti-intuitive notion of incompatibility without becoming inconsistent. This notion of incompatibility has profound repercussions on what this theory can possibly mean and these matters will of course be discussed extensively in the forthcoming chapters.

Projection operators

Closely related to the notion of the state vector and a basis $\{|i\rangle\}$ is the concept of a projector. A *projector* is an operator P that may act on vectors in a vector space like $\mathcal{H}$ and it projects the vectors along a particular axis, or in general on some subspace of $\mathcal{H}$. By virtue of this defining property applying a projector P twice on any vector gives the same result as applying it once: $P^2 = P$. Note that $\mathbf{1} - P$ is also a projection operator as it also squares to itself. We can rewrite this as $P(\mathbf{1} - P) = 0$ which amounts to saying that P and $\mathbf{1} - P$ project on orthogonal subspaces of $\mathcal{H}$. So given a projection operator one can make an orthogonal decomposition of the Hilbert space. On vectors in the first subspace the projector act as the unit operator, and on the vectors in the orthogonal complement it acts like the zero operator. This observation is highly relevant if one wants to assign properties to a quantum state. A projector P assigns a truth value to a state, but only if the state vector sits entirely in the subspace on which P projects, or its orthogonal complement. Clearly if the state vector has components in both, you cannot say it has the property nor can you say that it has not. But in that case there are other projection operators that do a better job, because there are always subspaces which contain that state vector or to which that vector is orthogonal. The notion of projectors plays an important role in the theory of quantum measurement as we will see in the next section.

Elementary projectors. One easily verifies that the projector P_j which projects on the axis corresponding to the basis vector $|j\rangle$ is given by:

$$P_j = |j\rangle\langle j|\,, \tag{II.2.8}$$

and indeed its square equals itself and applying it to a state vector and using (II.2.6) yields:

$$P_j\,|\Psi\rangle = \Sigma_i \alpha_i\,|j\rangle\langle j|i\rangle = \alpha_j|j\rangle\,,$$

which is exactly the component along the j-axis, i.e. $\langle j|\Psi\rangle\ |j\rangle$. Note that any sum over a subset of P_i is also a projection operator (because they mutually commute), and so is $|\Psi\rangle\langle\Psi|$ for any state $|\Psi\rangle$.

Consider 'bracketing' an elementary projector in some state:

$$p_i = \langle\Psi|\,P_i\,|\Psi\rangle = |\langle i|\Psi\rangle|^2| = |\alpha_i|^2\,, \tag{II.2.9}$$

it yields the component along the basis vector squared. This is the probability p_i of finding the particle in the state $|i\rangle$ in an appropriate measurement. The normalization condition (II.1.9) is nothing but the statement that the total probability of finding the system in some state equals one, as it should.

Completeness. One now can also understand that the set of elementary projection operators satisfies the so-called *completeness relation*, which amounts to the statement that

$$\sum_i |i\rangle\langle i| = \mathbf{1}\,. \tag{II.2.10}$$

This means that it works as the identity operator: acting on any state vector $|\psi\rangle$ it gives back the same state. The completeness relation is also referred to as the *projective decomposition of the identity operator*, since it is the operator equivalent of the statement that any state vector can be decomposed in its components with respect to some frame.

Observables and projectors. From the orthonormality relations of eigenvectors $\{|a_j\rangle\}$ of an observable A, and the properties of the corresponding elementary projectors P_j, one may show that we can actually write the operator A as:

$$A = \sum_j a_j P_j\,.$$

Needless to say that all projection operators are observables (as $P = P^\dagger$), but not the other way around!

Projectors on subspaces of $\mathcal{H}$. It is not hard to see that along these lines we can construct projectors that project

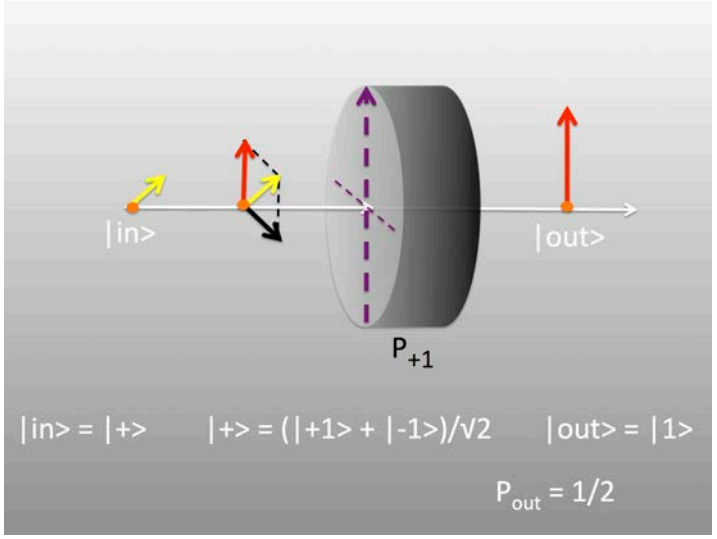


Figure II.2.6: *A photon polarizer.* A polarizer projects the photon onto a particular polarization state. There is a calculable probability for the photon to come through, after which it is fully polarized in the selected direction.

on a subspace of the Hilbert space, by adding up some subset Γ of elementary projectors:

$$P_\Gamma = \sum_{j\in\Gamma} P_j \,.$$

Such operators play an important role in the assignment of quantum properties to states in Hilbert space.

Photon polarizers are projectors. Photons can be projected on certain subspaces of the full Hilbert space, and these operations are quite familiar and dear to all of us. We can use a *color filter* to project on a certain subspace of wavelengths or frequencies. For example, you want to filter out the UV component of the light if you are high up in the mountains. But in the present context of qubits we should rather think of a *polarizer* which projects the polarization vector on a particular axis. As we have indicated in Figure II.2.6 the polarizer P_{+1} does actually more than just projecting the state, it projects the in-state $|+\rangle$ on the chosen $|+1\rangle$ *direction* of the polarizer, but then renormalizes the state to a vector of length one, so the outstate is $|+1\rangle$. The magnitude of the incoming component tells you the probability that the photon will be transmitted, so $p_{out} = (1/\sqrt{2})^2 = 1/2$. And that is what your fancy polaroid shades are really about. It is indeed a projector in the sense that if we let the photons that come through some polarizer, and subsequently let them go through an identical polarizer then all the photons will get through. If one rotates the second polarizer by 90 degrees, then that projects on the orthogonal subspace, and a photon that gets through the first polarizer will be blocked by the second. To check this you need two Ray-Bans, or if you are blessed with the curiosity of a true scientist you would happily break the one and only one you have in two pieces of course.

Note that for a large number of photons the result reproduces the classical result, if one identifies the reduction in the light intensity due to the polarizer with the ratio of the number of outgoing and the number of incoming photons. In the classical Maxwell theory, the light intensity is given by the square of the electric field. The classical field $\mathbf{E}$ is literally projected, giving the factor $1/\sqrt{2}$ in the magnitude of the projected component. And its square does give the reduction factor $1/2$., the same as in the quantum case. But again, for a single photon there is no classical description, and to explain the single photon experimental results one has to go quantum.

Raising and lowering operators

Let me try to make you more familiar with thinking about dynamical variables as operators or matrices by demonstrating a different use of the algebra of observables as operators on states. You may think of a system having some basic operator Q with its associated eigenvalues and eigenstates. We also require that the system has some ground state that we for the moment assume to be

a unique lowest state $|0\rangle$ with $Q|0\rangle = q_0|0\rangle$. Then we may search for operators $A^{\pm}$ that satisfy the relation:

$$[Q, A^{\pm}] = \pm q A^{\pm} . \qquad \text{(II.2.11)}$$

Writing this expression out we obtain the following property of the state $A^{\pm}|\psi_n\rangle$,

$$Q(A^{\pm}|\psi_n\rangle) = (q_n \pm q)(A^{\pm}|\psi_n\rangle) .$$

This means that starting with an eigenstate of Q, the operators $A^{\pm}$ create again an eigenstate of Q with a higher (lower) eigenvalue. Such *raising* and *lowering* operators are extremely useful because they would in principle allow you to create the excited states from the ground state; they allow you to move through the spectrum of Q eigenstates and are therefore also called *laddering* or *step* operators. Clearly the raising operators can be written in an explicit form as:

$$A^{+} = \Sigma_n |n+1\rangle\langle n| . \qquad \text{(II.2.12)}$$

Such a setup works only if the eigenvalues q_n are evenly spaced, in other words if $q_n = q_0 + nq$, but this is quite often the case.

Let us see how this works out for the example of the Hamiltonian $H_1 = Z$ of the previous subsection. The step operators are now the following linear combinations:

$$Z_+ = |1\rangle\langle -1| \;\Leftrightarrow Z_+ = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} , \qquad \text{(II.2.13)}$$

and

$$Z_- = (Z_+)^{\dagger} = |-1\rangle\langle 1| \;\Leftrightarrow Z_- = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} . \qquad \text{(II.2.14)}$$

They are not hermitian but, as advertised, they satisfy indeed the commutation relations (II.5.21) with $q = 2$, and they further more satisfy:

$$[Z_+, Z_-] = Z ,$$

which is just the Hamiltonian.

Now check that they step us through the spectrum of states. The ground state is in this case the state $|-1\rangle$ with lowest eigenvalue -1. Acting with the raising operator Z_+ yields:

$$Z_+|-1\rangle = |+1\rangle \Leftrightarrow \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} ,$$

with eigenvalue $+1$. You may want to check that the raising operator applied to the highest eigenstate $|+1\rangle$ yields zero and a similar statement holds about applying the lowering operator and the lowest energy or ground state.

We may turn the argument around and say that a lowering operator can be used to find the ground state $|\psi_0\rangle$ (up to some constant phase factor), by *requiring* $A_-|\psi_0\rangle = 0$ in the present case:

$$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 0 \\ \alpha \end{pmatrix} = \alpha \begin{pmatrix} 0 \\ 1 \end{pmatrix} \;\Rightarrow\; \alpha = 0 ,$$

from which follows that $|\psi_0\rangle = |-1\rangle$, up to the phase factor α.

The action of the step operators on the states is summarized in simple spectral diagram in Figure II.2.7. Note that the figure is also supposed to imply the fact that

$$Z_{\pm}|\pm 1\rangle = 0 ,$$

where the '0" on the right-hand side is the zero vector in the vector space. This zero does not represent a physical state as it has norm zero. The spectrum is bounded: it has a so-called highest and lowest weight state.

State operators. These operators and the pictures that represent their actions are quite useful in situations that are more complicated than qubits. What they allow you to do, is to give a different symbolic representation of the general qubit state (II.1.2), as we can write:

$$|\psi\rangle = (\alpha + \beta Z_+)|-1\rangle = \hat{\psi}_+|-1\rangle , \qquad \text{(II.2.15)}$$

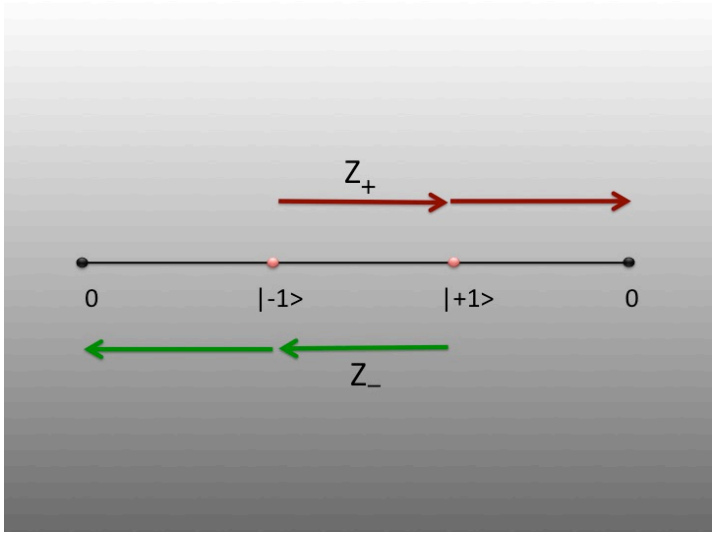


Figure II.2.7: *Step operators.* The action of the step operators $Z_\pm$ on the basis states $|\pm\rangle$. It is also implied that $Z_\pm|\pm 1\rangle = 0$, where 0 is the zero vector, which is not a physical state.

or alternatively:

$$|\psi\rangle = (\alpha Z_- + \beta)|+1\rangle = \hat{\psi}_-|+1\rangle\,. \qquad \text{(II.2.16)}$$

What this equation shows is that there is a correspondence between states and operators, if we know either a 'lowest weight' or a 'highest weight' reference state $|0\rangle_\pm$, defined by the conditions,

$$Z_-|0\rangle_- = 0 \text{ or } Z_+|0\rangle_+ = 0\,,$$

which, as we saw, yielded that the lowest or ground state is $|0\rangle_- = |-1\rangle$. What we learn is that there is an equivalence between specifying a state vector $|\psi\rangle$, and an operator $\hat{\psi}$ that acts on a given ground state $|0\rangle$. It is this perspective which turns out to be essential for understanding the spectrum of quantum particles and fields.

Figure II.2.8: Truth is in the eye of the beholder.*Time's eye* (1949) by Salvador Dali. (©Salvador Dalí, Fundación Gala-Salvador Dalí)

Quantum measurement

Physics as a science is deeply empirical. Theories have to be thoroughly tested by experiments and have to be adapted or refuted if they fail to be confirmed. Experiments involve measurements in which the features of the proposed theory are observed by some means. This means that quantum theory also features the subtle, if not exotic, concepts like the linear superposition principle and the possibility of entangled states. The basic theoretical features were hard to put to test at the time when the theory was formulated, because the experimental techniques were not sophisticated enough to reach the necessary degree of precision. The story of quantum measurement therefore has a rich history. The first dramatic pseudo experimental developments consisted of the well-known 'gedanken' or 'thought' experiments devised by none less than Einstein and Schrödinger themselves. Schrödinger's cat addressed the problematic side of the outrageous idea

that a cat could be in a state that is a linear combination of a 'dead' and an 'alive' state, and we discussed it in the previous chapter on page 266. The other is the EPR paradox, addressing the problematic aspect of non locality as a direct consequence of having spatially separated particles in an entangled state. This led to the view that quantum theory would be an incomplete theory to which 'hidden variables' would have to be added to make it local and causally consistent. It took fierce debates like the Einstein–Bohr debate, and it caused a search for alternative interpretations or even theories like the 'hidden variable' theory of David Bohm and the 'many worlds' interpretation proposed by Hugh Everett in 1958.

Our strategy in this book is that using our knowledge of states and observables as discussed so far, we present the commonly adopted (called orthodox by some) Copenhagen interpretation of measurement in this chapter, primarily because it has never been falsified, quite the opposite. Indeed it has been vindicated by numerous extremely refined recent experiments. Yet, not everybody is quite comfortable with the situation and we will get to some of the paradoxes and their (experimental) resolutions into more detail in Chapter II.4.

The question of quantum measurement has two parts to it: part one answers the question: given that the system is in a state $|\psi\rangle$ what can we say about the measurement outcome of some observable A. And the second part answers the question: how does a measurement affect the state $|\psi\rangle$? We will see that in quantum theory object and subject are, strictly speaking, no longer separable.

Probabilism. The interpretation of the wavefunction is at first sight quite bizarre: it is a measure for where the particle may be found if one is to make a measurement. More precisely, its square gives the probability density of finding the particle at position x at time t. Expressed in a compact formula it reads simply: $P(x,t) = |\Psi(x,t)|^2$. Probability? What? Didn't we completely specify the state and now at once we start talking about the odds of finding the particle somewhere. Is that all we can do? Can't we do better? Good question, so, let me quote what Richard Feynman said on this remarkable quantum state of affairs in part three of his famous Lectures on Physics.

> We would like to emphasize an important difference between classical and quantum mechanics. We have been talking about the probability that the electron will arrive in a given circumstance. We have implied that in an experimental arrangement (even in the best possible one) it would be impossible to predict exactly what would happen. We can only predict the odds! This would mean, if it were true, that physics has given up on the problem of trying to predict exactly what will happen in a given circumstance. Yes! Physics has given up. We do not know how to predict what would happen in a given circumstance, and we believe now that it is impossible – that the only thing that can be predicted is the probability of different events. It must be recognized that this is a retrenchment in our earlier ideal of understanding nature.
>
> *Richard Feynman, Lectures on Physics, Part III*

This quote characterizes the dramatic change of perspective on our capability to 'understand' the fundamental properties of nature. It was in fact the Austrian physicist Max Born who forcefully argued for this probabilistic interpretation of quantum mechanics, and he received the Nobel prize in 1935 for this work. This interpretation is usually referred to as the *Kopenhagener Deutung*, or *Copenhagen interpretation*, of quantum mechanics.

Classical versus quantum measurements. Measurement in classical physics is conceptually rather trivial: One simply observes the classical state variables with a finite precision and thereby approximates the variable as a real number with a finite number of digits. The accuracy of measurements is limited only by background noise and

the precision of the measuring instrument. The crucial assumption is that one can make any such measurement without changing the state of the system. This implies that the order in which one makes measurements is irrelevant, and therefore there is no restriction on which variables could be measured 'simultaneously.'

In the quantum setup we describe a particle with a wavefunction which may be spread out over all of space. The fact that the wavefunction is spread over all of space, however, does not mean that the particle is at many places simultaneously, or that we could observe it in different places at the same time. It does not even mean that the particle is actually in some definite place and that we only happen to just not know *where* it is. The particle state is a probability amplitude, referring not to the probability where the particle actually *is* but to where it might be found upon making a position measurement. As we will see it basically doesn't make sense to talk about *where* the particle is before we observe it. In general the wavefunction tells us that the particle *is*, rather than where it is.

Indeed, that situation is quite different from the proposition that we know someone is in a room behind a closed door, and we do not know where in the room this person exactly is, because in that case we know for sure that the person will be definitely somewhere and we may assign a certain probability distribution as to where she is. That distribution however reflects *our* ignorance, *our* not-knowing the exact state. It describes our lack of knowledge as observer, not the actual state this person is in.

In quantum theory a given extended wavefunction specifies the state of the particle *completely*, and knowledge of that state does not allow us to deduce where the particle is; its position is just not determined, in that state *it has no position a priori* and it therefore makes no (quantum) sense to talk about it! The fundamental difference between a possible classical probability which reflects our lack of knowledge about the system, and the inescapable uncertainty that occurs even if we know the state exactly is that the quantum probability refers to an intrinsic property of the system and not to the state of knowledge that an observer like you or me might or might not have about that system. Yet, at the same time, the state limits fundamentally what an observer could possibly get to know about the system. As a consequence the measurement process in quantum mechanics is not at all trivial.

Leaving a trace. A misleading aspect of measurement theory is that the term measurement suggests that it is necessary to have an experimenter who is handling some intricate device to collect data. This is not the case. As a matter of principle, it only matters that the system interacted with something, somewhere, at some time, and that that interaction affected the state of the system. The interaction may have left a trace somewhere, an indelible mark, without any experimenter caring about it or even being aware of it. In that sense the notion of measurement is much more abstract, and less anthropocentric than you might have thought. It is like 'forensic science,' where one is searching for traces of past interactions call it of 'measurements' – that took place a long time ago: finger prints, car keys, or sunglasses left on a table, or phone calls, and photographs left on a remote server. A measurement is anything that leaves some discernible trace somewhere, at some instant in time.
So if I engage into an interaction with a particle, its behavior may have been influenced by previous interactions I have no knowledge about, and that may in turn lead to unexpected outcomes in my experiment. Something I better be aware of. It is the hidden constraints that often present an invisible yet fatal flaw. We return to these questions in Chapter II.4. □

Another notable difference with classical mechanics is that in many instances the set of observable states is discrete, with quantized values for the physical variable. It is this property that has given the theory of quantum mechanics its name.

Maybe the most profound difference is that quantum measurement typically causes a radical alteration of the state vector. Before the measurement of an observable we can only describe the possible outcomes in terms of probabilities, whereas after the measurement the outcome is known with certainty, and the wavefunction is irrevocably altered to reflect this. In the Copenhagen interpretation of quantum mechanics the wavefunction is said to 'collapse' when a measurement is made.

In spite of the fact that quantum mechanics makes spectacularly successful predictions, the fact that quantum measurements are inherently probabilistic and can 'instantly' alter the state of the system in such a disruptive manner has caused a great deal of confusion and controversy. In fact, one can argue that historically the field of quantum computation emerged from thinking carefully about the measurement problem.

No cloning!

If measuring a quantum state changes it, you may wonder whether it is not a smart idea to copy such a state, before making the measurement. Take one and make two identical ones out of it by using a quantum Xerox machine. The answer is simply that this just cannot be done. Quantum copying is a no-go! This exceptional feature create the possibility of a novel type of 'quantum security:' Information that cannot be copied without destroying it. This makes the no-cloning principle a blessing in disguise.

What I am trying to tell you is that reading a quantum book will change it in unpredictable ways. You might actually want to avoid trouble with the librarian by copying the quantum book before reading it. But even this precautionary measure is obstructed by a quantum *no cloning theorem*, which was first formulated by William Wootters and Wojciech Zurek and by Dennis Dieks in 1982.

Suppose I have one particle in a particular state, and I want to bring another particle into exactly the same state. Then I have to look at the state of particle one in order to know what state to bring particle two in. But, by doing so, I have to affect the state of particle one. The best I can do in general is to bring particle two in the state particle one was in before, but then particle one is no longer in that state. This remarkable property can be shown to hold rigorously: quantum states cannot be copied, but they may be transferred from one system to another. And thinking in terms of securing information and beating our National Security Agencies with respect to protecting our privacy, this no-cloning may turn out to be a blessing in disguise. And it is.

More precisely, the no-cloning theorem amounts to the statement that for an arbitrary state $|\psi_1\rangle$ on one qubit and some particular state $|\phi\rangle$ on another, there is no quantum device $[A]$ that transforms $|\phi\rangle \otimes |\psi_1\rangle \rightarrow |\psi_1\rangle \otimes |\psi_1\rangle$, i.e. that transforms $|\phi\rangle$ into $|\psi_1\rangle$, while leaving the old $|\psi_1\rangle$ unaffected. If U_A is the unitary operator representing A, this can be rewritten $|\psi_1\rangle|\psi_1\rangle = U_A|\phi\rangle|\psi_1\rangle$. For a true cloning device this property has to hold for any other state $|\psi_2\rangle$ as well, and we must also have $|\psi_2\rangle|\psi_2\rangle = U_A|\phi\rangle|\psi_2\rangle$. It is not hard to demonstrate that the existence of such a device leads to a contradiction. Since $\langle\phi|\phi\rangle = 1$ and $U_A^\dagger U_A = 1$, the existence of a device that can clone both ψ_1 and ψ_2 would imply that

$$\begin{aligned}\langle\psi_1|\psi_2\rangle &= (\langle\psi_1|\langle\phi|)\,(|\phi\rangle|\psi_2\rangle) \\ &= (\langle\psi_1|\langle\phi|U_A^\dagger)\,(U_A|\phi\rangle|\psi_2\rangle) \\ &= (\langle\psi_1|\langle\psi_1|)(|\psi_2\rangle|\psi_2\rangle) = \langle\psi_1|\psi_2\rangle^2 .\end{aligned}$$

The property $\langle\psi_1|\psi_2\rangle = \langle\psi_1|\psi_2\rangle^2$ only holds if ψ_1 and ψ_2 are either orthogonal or aligned meaning that either $\langle\psi_1|\psi_2\rangle = 0$ or 1. It does not hold for arbitrary values of ψ_1 and ψ_2, so there can be no such general purpose cloning device. In fact, in view of the uncertainty of quantum measurements, the no-cloning theorem does not come as a surprise. If it were possible to clone wavefunctions, it would be possible to circumvent the uncertainty of quantum measurements by making a large number of copies of a wavefunction, measuring different properties of each copy, and reconstructing the exact state of the original wavefunction.

The probabilistic outcome of measurements

In the formalism of quantum mechanics the possible measurement outcomes of an observable quantity A are given by the eigenvalues of the matrix A. For example, the three Pauli matrices, defined in equation (II.2.2), all have the same two eigenvalues $\lambda_\pm = \pm 1$. This means that the possible outcomes of a measurement of the spin *in any direction* can only be plus or minus one. This is fundamentally different from a spinning object in classical physics, which can spin at any possible rate in any direction. The observed value of any component of a classical spin in this picture could be any real number between -1 and $+1$. This confirms that quantum mechanics is counter-intuitive and subtle indeed.

If a quantum system is in an eigenstate of an observable, then the outcome of measurements of that observable is 100% certain. For example, imagine we have a qubit in the state with $\alpha = 1$ and $\beta = 0$, so that $|\psi\rangle = |+1\rangle$. It is then in the eigenstate of Z with eigenvalue $z = +1$ and the measurement of Z will always yield that value. This is depicted in Figure II.2.9(a), and is reflected in the mathematical machinery of quantum mechanics by the fact that for the spin or polarization operator in the $z-$direction, $A = Z$, the eigenvector with eigenvalue $\lambda_+ = +1$ is $|+1\rangle$ and the eigenvector with $\lambda_- = -1$ is $|-1\rangle$. In contrast, if we make measurements in another direction, e.g. $A = X$, the outcomes become probabilistic. The outcome is still $+1$ or -1, but there are calculable probabilities for each value to occur. So the take-away message here is that it is not the values of possible outcomes that change, only the probability by which they will occur. Quantum theory is dealing with 'certain uncertainties', so to say. This is depicted in Figure II.2.9(d). The eigenvectors of X are:

$$|+\rangle = \sqrt{\frac{1}{2}}(|+1\rangle + |-1\rangle) \quad \text{and} \quad |-\rangle = \sqrt{\frac{1}{2}}(|1\rangle - |-1\rangle)\,.$$

In general the probability of finding the system in a given state through a measurement is computed by first writing the given state $|\psi\rangle$ as a linear combination of the eigenstates $|a_k\rangle$ of the matrix A corresponding to the observable, i.e.

$$|\psi\rangle = \sum_k \beta_k |a_k\rangle \text{ with } \beta_k = \langle a_k|\psi\rangle\,.$$

The notation $\langle a_k|\psi\rangle$ means that the component β_k is indeed equal to the projection of the state vector $|\psi\rangle$ on the eigenvector $|a_k\rangle$. The probability of measuring the system in the state corresponding to eigenvalue a_k is then given by

$$p_k = |\beta_k|^2 = |\langle a_k|\psi\rangle|^2\,. \qquad \text{(II.2.17)}$$

As we discussed briefly before, this is why the coefficients β_k in the expansion of the state $|\psi\rangle$ in a set of eigenstates of some observable are called *probability amplitudes*, amplitudes because it is only after squaring them that one obtains the probabilities for a certain measurement outcome. And the normalization condition on the state vector is just the statement that the total probability to find the system in one of the allowed states, equals one. The other two pictures of Figure II.2.9 give smilar distributions for an incoming $|+\rangle$ state. In Figure II.2.10 we given the corresponding distributions of electrons hitting the screen perpendicular to the beam. This is what one sees preparing the beam

in the incoming state and then measuring its polarization along some given axis.

So, what constitutes a measurement? I have been somewhat cavalier in talking about the notion of a measurement, while showing you nice and clean figures of some idealized experiments. Indeed at this stage, where we for example talk about spin polarization measurements, we have a situation in mind where we distinguish three stages in a measurement experiment.

(i) A preparatory stage, where we prepare the particle(s) so that the spin is in the desired state. For example we have electrons coming in and by using a Stern–Gerlach device (this will be explained in the next chapter) we can split the beam into two with opposite polarizations along an axis one may choose. This way one may prepare a beam of spins in some definite and identical polarization state up to an overall phase.
(ii) A first stage of the measurement, where we let the prepared beam sequentially interact with some other devices, which make up the experiment.
(iii) The second and final stage of the measurement, where we actually have a 'screen' or other counting device. So, in the end we measure a probability distribution that can be compared with a theoretical prediction, and potentially falsify our theory.

The purist may say that only the very last stage constitutes the measurement, so where the distribution over the sample spaces of some pre-chosen set of observables is obtained by projecting the outcoming particle states.

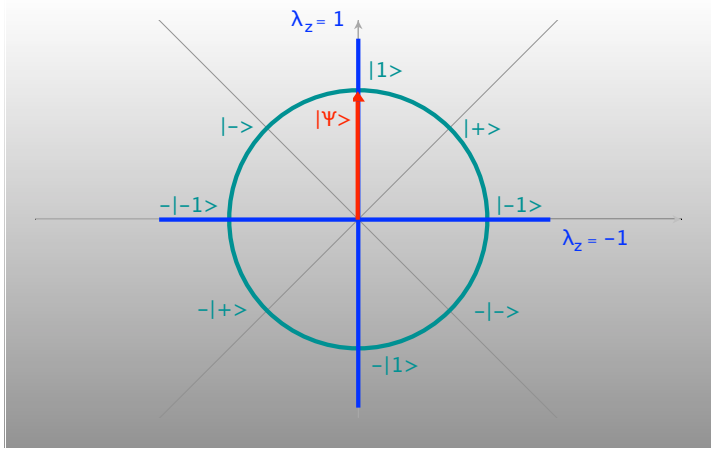
The projection postulate

> In classical physics, science started from the belief – or should one say, from the illusion? – that we could describe the world, or least parts of the world, without any reference to ourselves.
>
> *Werner Heisenberg*

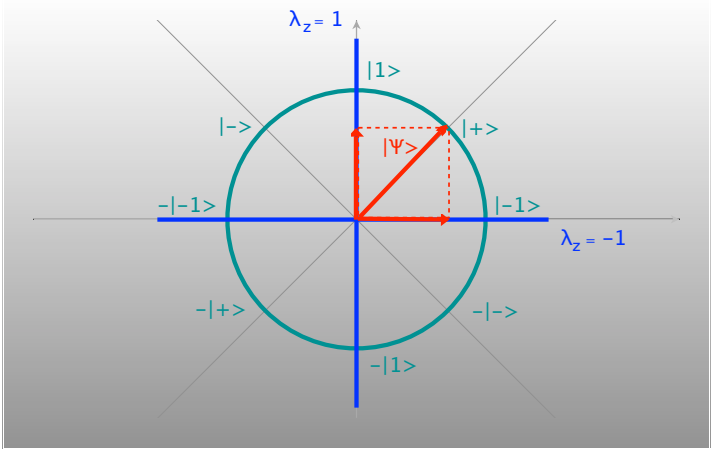
Apart from the probabilistic nature of measurement outcomes, a second remarkable aspect of quantum measurement is the fact that the act of making a measurement will generically change the state of the system. It is disruptive and will cause what is known as a '*collapse of the wavefunction.*' The mechanism is also known as the *projection postulate*, which was formulated by John von Neumann in the early days of quantum mechanics. This postulate is at this point an extra and in fact ad hoc postulate. Ad hoc, because the measurement process itself is just a quantum process and therefore should be completely described within the framework of the theory. The outcome should be 'calculable' from first principles and cannot be decreed by an additional postulate. In the end it is to be decided by ever more precise measurements whether or to what extent the postulate really holds and correctly represents all possible choices. But even then, the postulate including its range of validity should be 'proven' from first principles.
This being said, the reason this is so hard is because a typical realistic measurement device is a macroscopic, classical machine. So what I just said will be extremely complicated, because you have to model the effective interaction between quantum and classical degrees of freedom, basically by going all the way down to the quantum level in describing the apparatus.
In an operational sense the projection postulate so far has been confirmed by basically all experiments dedicated to test it. It is this 'success' which causes that the terminology and related picture of the measurement process persist in the mindset of most quantum practitioners .

Over the last few decades, physicists like to distinguish so-called *strong* and *weak* measurements. Let us comment on them subsequently.

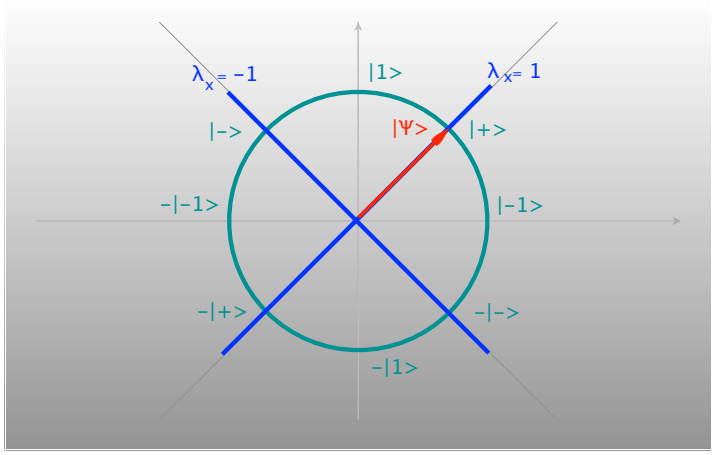
Strong measurements. The strong measurements are the most common ones. One observes a particular eigenvalue as we discussed, and the system makes then a transition exactly to the corresponding eigenstate. This type of measurement does confirm the postulate by definition.



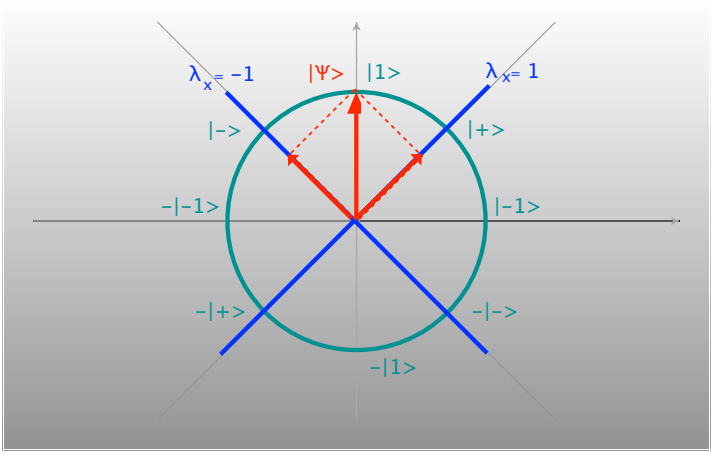
(a) Measurement spin polarization along z-axis, of the state $|\psi\rangle = |1\rangle$. Outcome: probability $p_z(+1) = 1$ and $p_z(-1) = 0$.



(b) Measurement spin polarization along z-axis, of the state $|\psi\rangle = |+\rangle$. Outcome: $p_z(+1) = p_z(-1) = 1/2$.

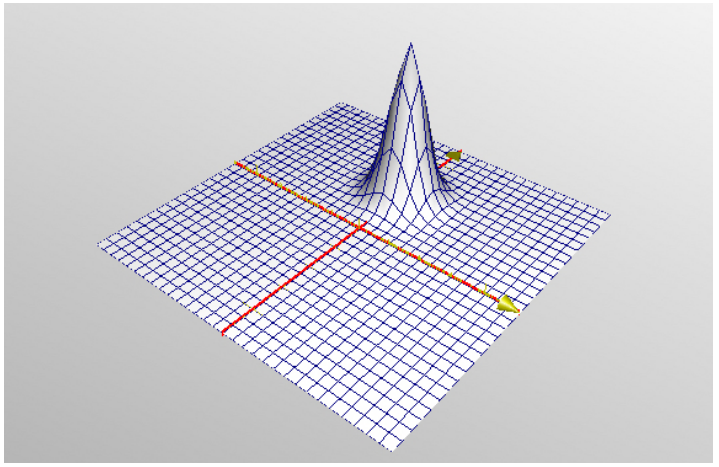


(c) Measurement spin polarization along x-axis, of the state $|\psi\rangle = |+\rangle$. Outcome: $p_x(+1) = 1$ and $p_x(-1) = 0$.

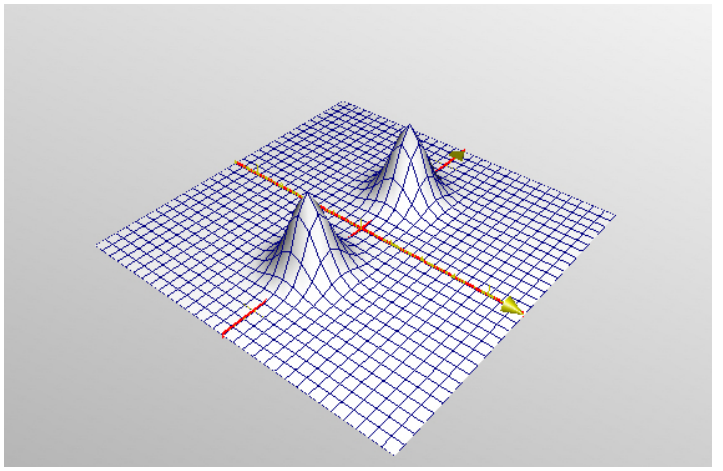


(d) Measurement spin polarization along x-axis, of the state $|\psi\rangle = |1\rangle$. Outcome $p_x(+1) = p_x(-1) = 1/2$.

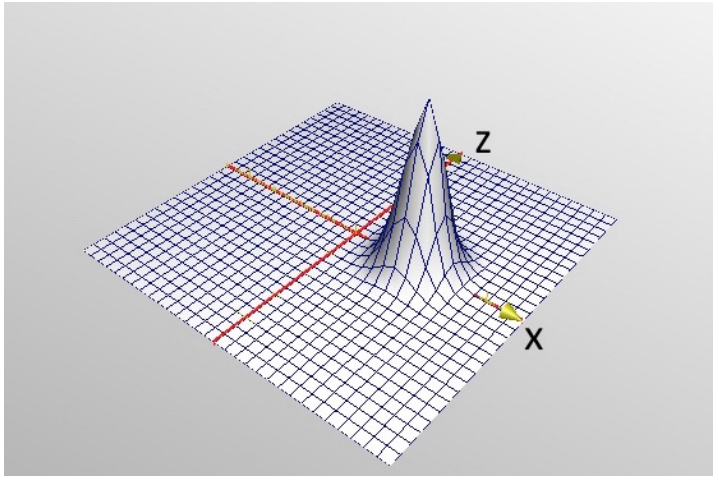
Figure II.2.9: *Spin polarizations.* Graphical representation of spin polarization along different axes. The projections of the red state vector $|\psi\rangle$ along the axes of the measurement frames gives the probability amplitude for the outcome to be plus or minus one.



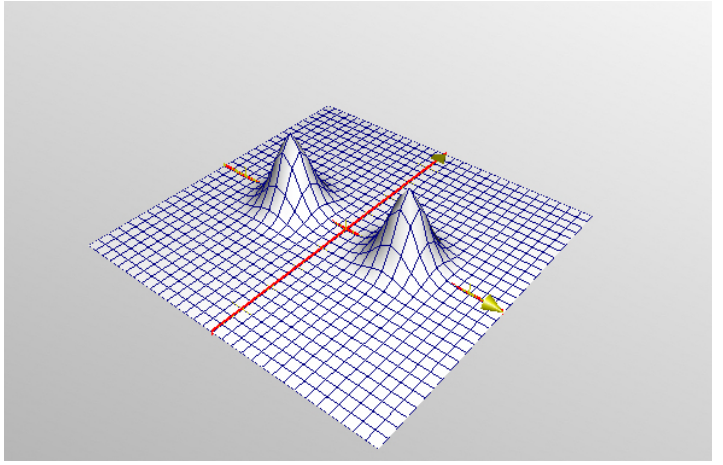
(a) Measurement spin polarization along z-axis, of the state $|\psi\rangle = |1\rangle$. Outcome: probability $p_z(+1) = 1$ and $p_z(-1) = 0$.



(b) Measurement spin polarization along z-axis, of the state $|\psi\rangle = |+\rangle$. Outcome: $p_z(+1) = p_z(-1) = 1/2$.



(c) Measurement spin polarization along x-axis, of the state $|\psi\rangle = |+\rangle$. Outcome: $p_x(+1) = 1$ and $p_x(-1) = 0$.



(d) Measurement spin polarization along x-axis, of the state $|\psi\rangle = |1\rangle$. Outcome $p_x(+1) = p_x(-1) = 1/2$.

Figure II.2.10: *Spin polarization measurements.* We have visualized the probability distributions discussed in the previous figure, in counts on a $z - x$ screen. The incoming beam is coming down along the y-axis after passing through a polarizing beamsplitter. The width of the distribution is supposed to reflect the width of the beams.

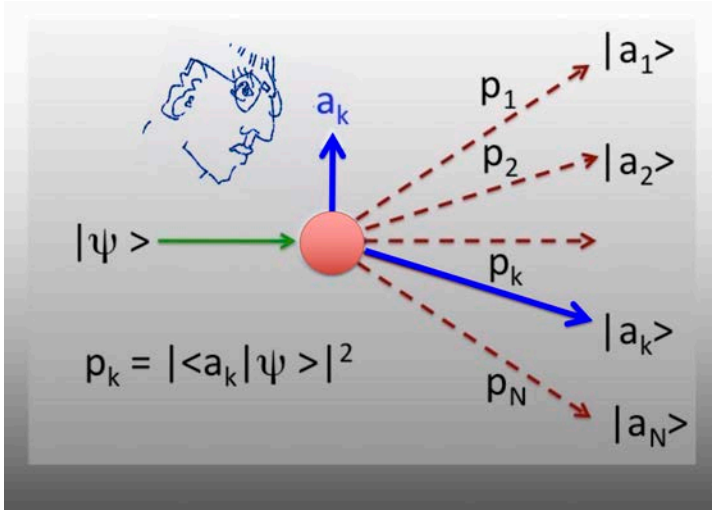


Figure II.2.11: *Projective measurement.* For the incoming state $|\psi\rangle$ there is a probability p_k, equal to the projection of the state on the eigenvector squared, to observe the (eigen)value a_k.

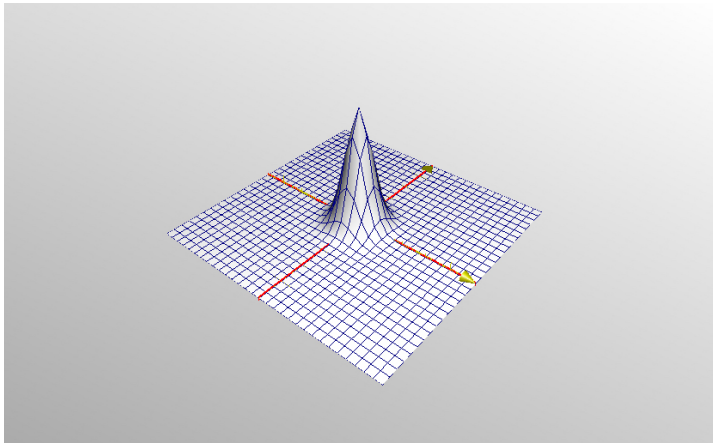
What happens is depicted schematically in Figure II.2.11. We start with a system in some state $|\psi\rangle$ and we make a measurement of the observable A and find a value a_k, then the act of making the measurement changes the state $|\psi\rangle$ to the state $|a_k\rangle$, the eigenstate of A with observed eigenvalue a_k. What this means is that if we would act with A again immediately after, we would measure that same eigenvalue with 100% probability, and that seems like a reasonable thing to expect.

Weak measurements. Fortunately, one is of course free to invent whatever smart measurement schemes one wants to pursue, in order to – in a more subtle way – extract more information than the projection postulate would allow you to. This has lead to an interesting debate within the physics community about so-called *weak measurements* and *weak values*.

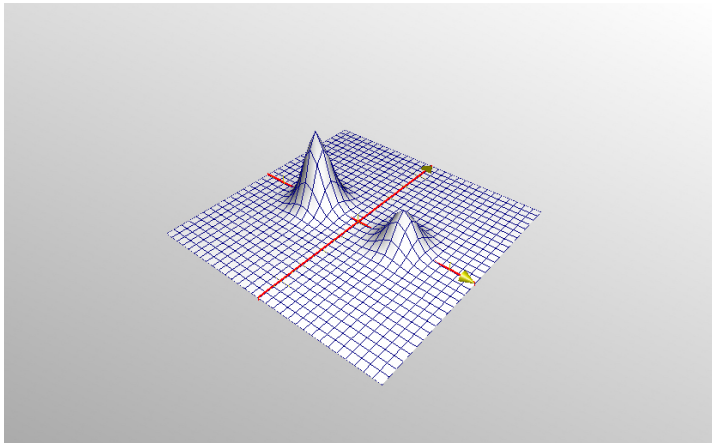
The idea is to make measurements where the interaction with the system is sufficiently weak so that it does not affect the incoming state. Yet, there is the possibility to observe a 'weak value' which would tell us 'something extra' about the state of system. As the state hasn't changed after the weak measurement, a strong measurement of another incompatible observable, made right after the weak one would not be affected. You should think of this as the subtle changes in the screen patterns of Figure II.2.12, like a small displacement in one of the peaks.

We have seen that a projective measurement with its collapse of the wavefunction amounts to a major disruption of the system, and here we consider the possibility to perturb the system in a subtle way, meaning weakly. These weak measurements may tell us something about the state of the system without really making a complete projection. In Figure II.2.12 we have depicted a scheme proposed by Aharonov, Albert and Vaidman, and show what happens to the particle distributions after we do such a weak measurement. We have incoming particles in a state $|\psi\rangle = (|+\rangle + \sqrt{2}|-\rangle)/\sqrt{3}$. In Figure II.2.12(a) we have the incoming beam and do no polarization measurement. In the second Figure II.2.12(b) we measure the polarization along the x-axis, and we see the expected splitting, with outcome $p_x(+1) = 1/3$ and $p_x(-1) = 2/3$. In Figure II.2.12(c) we start with an incomplete polarization measurement along the z-direction, which means that we apply a weak field so that the beam does not really split. This amounts to a small perturbation of the incoming beam. However, if directly after the weak measurement, we measure the x-polarization of the perturbed beam we observe a small displacement of the weak peak in the z direction as indicated in Figure II.2.12(d). The projection along the x-axis, however, takes place as usual, but one has succeeded in getting some extra information on the 'incompatible' z-polarization. It is this tiny shift in the z direction which amounts to the measurement of a *weak value*.

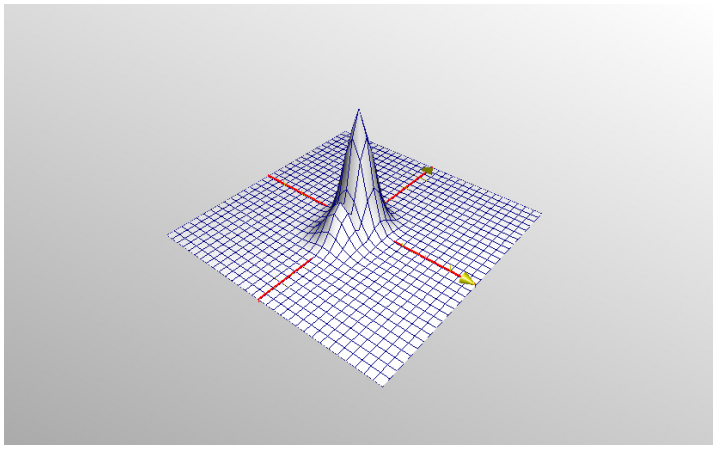
So here we have an example that illustrates the subtlety of the notion of measurement, the clue being that we have concocted a setup where we go beyond a simple projective measurement. It underscores that all interactions in some



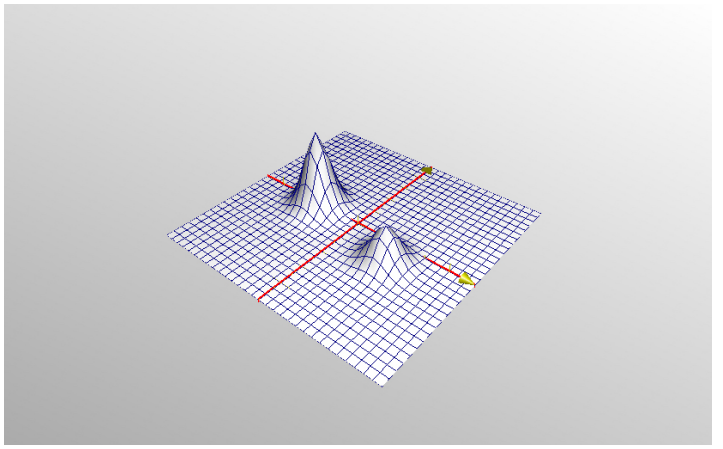
(a) Measurement spin polarization of the state $|\psi\rangle = (|+\rangle + \sqrt{2}|-\rangle)/\sqrt{3}$. Polarizers are turned off.



(b) Measurement spin polarization along x-axis, of the state. Outcome: $p_x(+1) = 1/3, p_x(-1) = 2/3$.



(c) A weak measurement of the spin polarization along z-axis, of the same state, yields a perturbed state.



(d) Measurement spin polarization along x-axis, of the perturbed state. Outcome $p_x(+1) = 2/3$ and $p_x(-1) = 1/3$. However the small peak is slightly shifted.

Figure II.2.12: *A weak spin measurement.* The incoming beam is coming down along the y-axis after passing through a z- and/or x-polarizing beamsplitter. The width of the distributions reflects the width of the beams. The results are explained in the text.

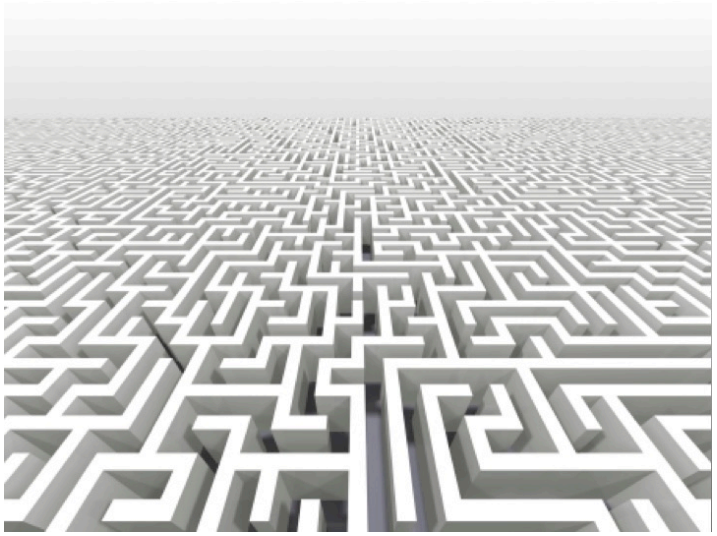


Figure II.2.13: *Logic and syntax.* In search of semantics?

way could be called a measurement, if you are willing to stretch the semantics of the term measurement.

Quantum grammar: Logic and Syntax

In the classical situation we speak of the phase space *of a system, to be contrasted with the* Hilbert space *for quantum systems. The fundamentally different structure of these two spaces has profound consequences for the logical and deductive structure of these theories. Whereas in the classical case properties of the system generally can be associated with subspaces of the total phase space, one has on the quantum level to distinguish the space of observables from the Hilbert space, and choose from possible consistent* frameworks *which are more restrictive. Within a framework certain* ***properties*** *can be unambiguously assigned, and deductive logic can be applied. This is illustrated for the cases of a qubit and a particle.*

Compatible observables allow for joint eigenstates and thus for those states one may assign a point in the joint sample space. A maximal subset of independent observables that are mutually compatible defines a *consistent framework* $\mathcal{F}$ to describe the system with. With the framework comes a sampling space $\mathcal{S}$ which is a kind of quantum equivalent of the classical phase space. So for the qubit example this is clear. A consistent framework could correspond to the Z observable, and we may describe all states of the qubit, as (normalized) linear combinations of the basis states $|\pm 1\rangle$ which are the eigenvectors of the Z observable as it makes up the framework.

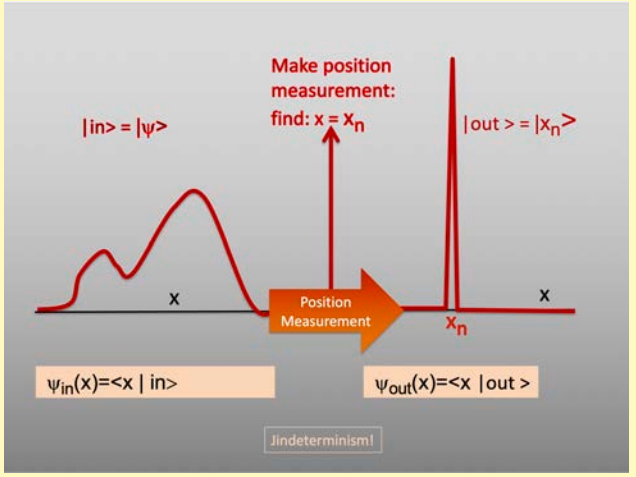
The framework for a quantum system is not unique, and the choice of framework depends on what question one wants to address and what aspect of the system one wants to study. If you make position measurements you use the Z-framework, and if you make momentum measurements you choose the X-framework. Let me emphasize however that a quantessence here is that there are observables which are not compatible with the framework. Logically speaking what this implies is that the observables incompatible with the particular framework you are using cannot be assigned a *meaning*. They are *meaningless* in that framework because there is no logical way one can decide whether a property referring to the values of incompatible observables is true or not. Henceforth quantum theory has well-defined observables that have the unusual feature that they cannot be part of a logically sound deductive argument within a given framework. Let us take a closer look at this statement and find out what this means for a classical particle and its quantum descendent.

Collapse of the wavefunction. In figure (a) below we give a graphic impression of what is called 'the collapse of the wavefunction.' If you think of the wavefunction as a probability amplitude, it makes actually a lot of sense, because you would expect that repeating the same measurement immediately after you have made the observation a_k would give exactly the same outcome with 100% certainty. But that can only be the case if the state has changed to the corresponding eigenstate $|a_k\rangle$ as decreed by the projection postulate. So the term 'collapse of the wavefunction' suggests that there is a violent physical action at a distance going on if we make a measurement, but that is totally misleading. The wavefunction which indeed encodes all there is to know about the state of the system represents a probability amplitude, and making a measurement can drastically change the probability of future measurement outcomes.

This is a familiar phenomenon. If I know that you are somewhere in town, I may have a rather uniform probability distribution for where you are that stretches all the way to the outskirts of the city. If you then suddenly happen to walk into my office, my probability distribution will indeed instantaneously collapse to some narrow spike that peaks right in front of my desk. But that doesn't mean that something is physically changing on the outskirts of town, nor will you be affected.

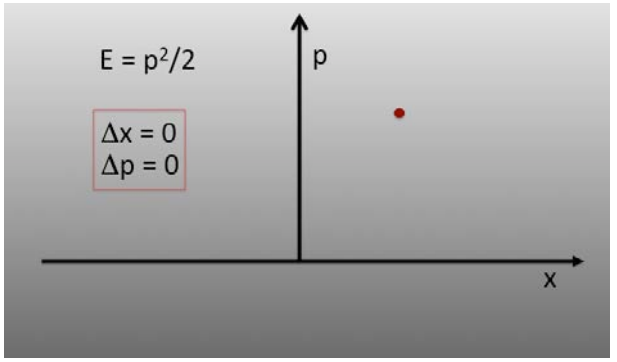
The quantessential difference between the quantum case and you is of course that the distribution I had in my mind about you was certainly *not* all there was to know about the system called 'you!'. It had more to say about my state of ignorance than about you. The measurement did not affect you nor places where you could have been. Apparently in quantum theory the strict separation of object and subject that reigns in classical physics is no longer valid: no longer any neutral observers, no peeking, or looking without touching.

In the classical context, the separation of object and subject is based on the assumption that it is in principle possible to make the effect of the measurement on the system arbitrarily small. This is no longer true in quantum theory.

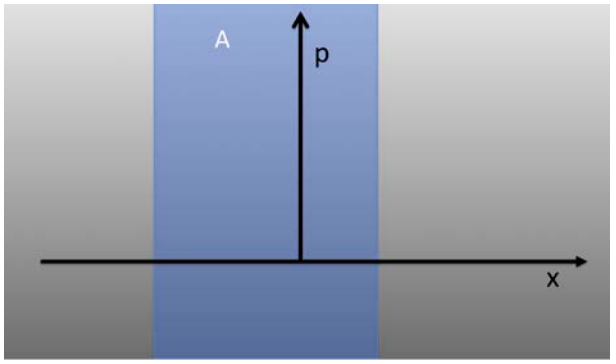


(a): *Collapse of the wavefunction.* A state $|\psi\rangle$ comes in and a measurement of the observable A is made. This yields with a probability p_n the outcome $x_n \in x$, and the state $|\psi\rangle$ instantly 'collapses' to the state $|x_n\rangle$.

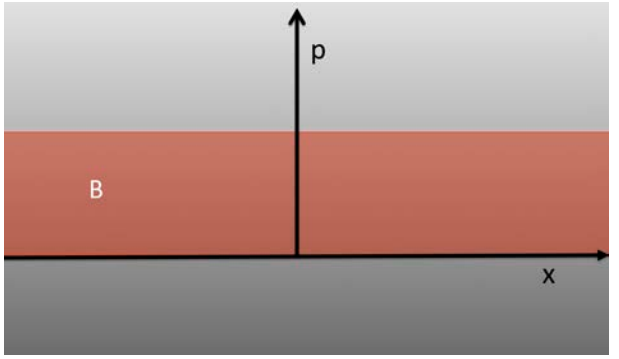
Sure enough, given a particular state there may be an appropriately chosen measurement that does not change the state, but in general it does change the state. So imagine how strange it would be if, after you read that quantum book, it changed. Never a dull moment, but alas nobody could guarantee you that the book would still make sense after you read it. A recipe for great applications in social media I think.□



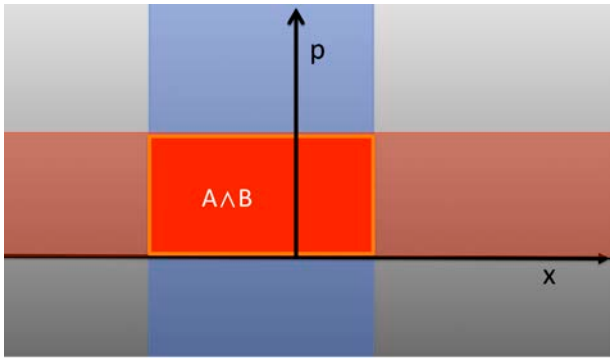
(b) Classically the state of a particle in one dimension is defined by its position x and momentum p , which define a point in its phase space $\mathcal{F}_{\mathrm{ph}}$.



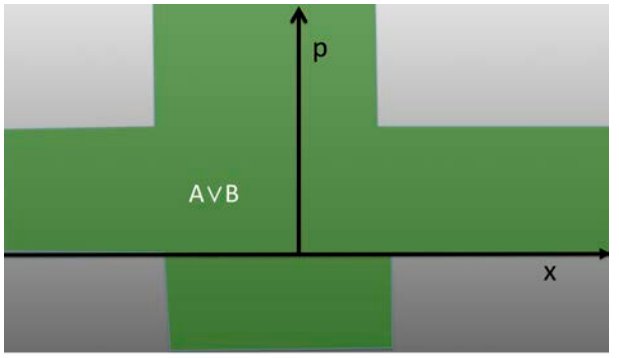
(c) The region corresponding to the proposition A: $\mathrm{x}_0 < \mathrm{x} < \mathrm{x}_1$ is shaded blue. It is true for a state if the point representing that state is in the blue region.



(d) The proposition B: $0 < \mathrm{p} < \mathrm{p}_1$ is true for all points in the dark red shaded region.



(e) The conjunction 'and' denoted as $\mathrm{A} \wedge \mathrm{B}$ corresponds to the bright red region.



(f) The conjunction 'or' denoted as $\mathrm{A} \vee \mathrm{B}$ corresponds to the green region.

Figure II.2.14: *Propositions in classical physics.* Propositions about the the position x and momentum p of a particle in one dimension and their conjunctions.

The case of a classical particle

Position and momentum are the basic observables that label the dynamical state of a particle which corresponds to a point in the phase space of the particle as illustrated in Figure II.2.14(b). These are basic because in the Newtonian 'framework' one has to specify the momentum and position at some initial time. Then the states at any other time would be determined provided we know the force acting on the particle. The fact that the momentum and position variables are basic also implies that other dynamical variables like energy can be expressed in them.

We can make propositions involving properties of particular states of the particle and find a yes/no answer to whether that proposition is true or false. Not only can we answer questions about the elementary properties but also about conjunctions of those. For example, we may ask whether a state has the property A: $\mathrm{x}_0 \leq \mathrm{x} \leq \mathrm{x}_1$. Then for all points x, p in phase space in the blue shaded region of Figure II.2.14(c) the answer is yes, and outside that region it would be no.

So we can assign a truth value '1' or '0' to the proposition A accordingly. Similarly we may ask for the p value to satisfy $0 \leq \mathrm{p} \leq \mathrm{p}_1$ and define it as proposition B, and then we get the picture of Figure II.2.14(d). Now we can ask for combined properties of x and p. For example, if may ask whether the property $\mathrm{A} \wedge \mathrm{B}$ (A **and** B) is true or not. The truth value of this conjunction can be calculated, and for the case at hand it equals the product of the truth values of A and B. This assignment requires of course that $\mathrm{AB} = \mathrm{BA}$, which means that the point has to be located in the bright red shaded rectangle as indicated in Figure II.2.14(e), the region that is the intersection of the shaded regions in the two previous figures.

Similarly, one may ask whether $\mathrm{x}_0 \leq \mathrm{x} \leq \mathrm{x}_1$ **or** $0 \leq \mathrm{p} \leq \mathrm{p}_1$ is valid, which means that we ask whether the property $\mathrm{A} \vee \mathrm{B}$ is true or not. This proposition is in the picture represented as the union of the shaded areas, which is the green shaded area in Figure II.2.14(f). Formally the truth value can be calculated by the formula $\mathrm{A} + \mathrm{B} - \mathrm{AB}$. The figures can be summarized in a conventional truth table as shown above, exactly as they are used in elementary (propositional) logic. So to find the properties of the classical particle, the physicists infer these from the rules of a simple deductive logical scheme that is mathematically represented by a Boolean algebra with variables that can only take two values, zero (false) or one (true).

Table II.2.1: Truth table for the propositions made in Figure II.2.14

A	B	A$\wedge$ B	A$\vee$ B
0	0	0	0
1	0	0	1
0	1	0	1
1	1	1	1

The case of a quantum particle

Let us now sketch what happens to the particle in the quantum arena. There is again a basic set of quantum observables 'X' and 'P'. And again one may ask at any moment what the value of any of the observables is and verify by measurement whether the proposition is true or false.

Sampling spaces. Here we first have to address the question of what the *sampling spaces* for these observables are. Let us allow two possibilities for the space in which the particle moves: it could be infinite and correspond to a straight line or it could be finite, say, a circle. The possible outcomes of position measurements would of course correspond to points in these spaces, meaning that the sam-

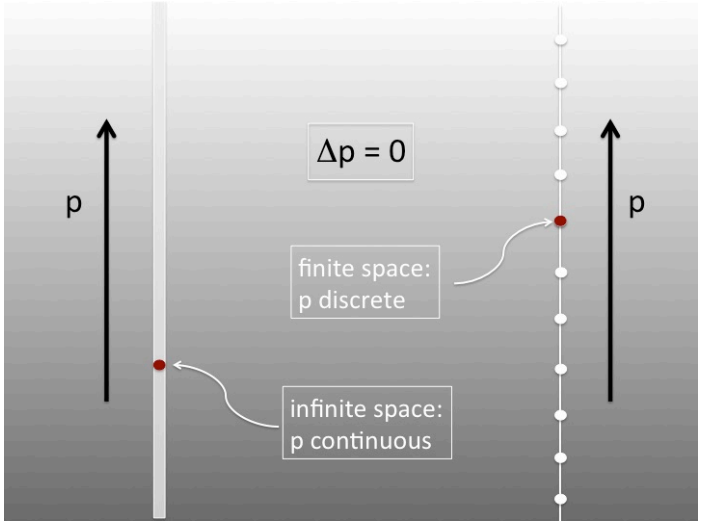


Figure II.2.15: *Sample space of momentum.* The sample space $\mathcal{S}_p$ for the momentum observable in the quantum case depends on the topology of the (continuous) configuration space $\mathcal{X}$ in which the particle moves.

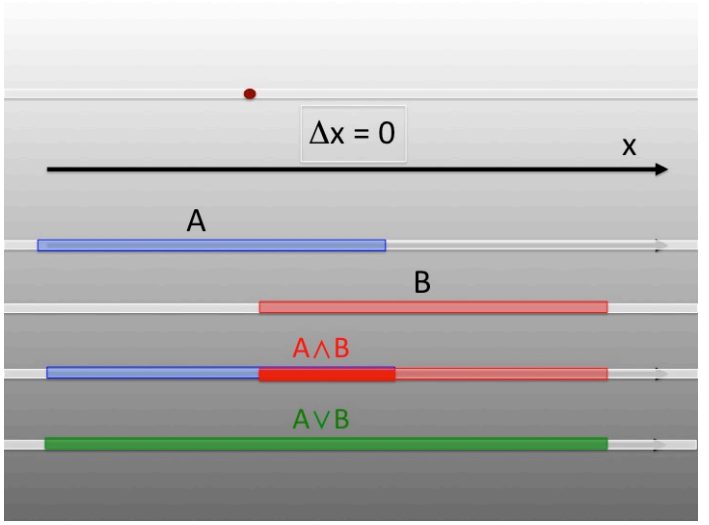


Figure II.2.16: *Sample space of position.* The sample space for the position observable is the real line. We indicated two possible propositions A and B, and their conjunctions.

ple space $\mathcal{S}_x \simeq \mathcal{X}$. However, the sample space for the momentum observable turns out to depend on the topology of the underlying configuration space.
If the particle lives on the real line and $\mathcal{X} \simeq \mathbb{R}$, then the possible values for the momentum variable are continuous just like the position variable, $\mathcal{S}_p \simeq \mathbb{R}$.
On the other hand, if the configuration space would be a circle $\mathcal{X} \simeq S^1$, then, as Bohr told us, the spectrum of the momentum becomes discrete and would in fact correspond to the set of integers denoted by $\mathcal{S}_p \simeq \mathbb{Z}$. We will treat the case of a particle on a circle in detail in the Chapter II.5. We have indicated the two possibilities in Figure II.2.15.

There is a third possibility here, that at first may strike you as utterly pointless but turns out to be quantessential and should not be overlooked. Imagine that the position space itself is *discrete and infinite*, like a one-dimensional lattice $\mathbb{Z}$, then, one should expect to find that the sample space for the momentum becomes a circle, $\mathcal{S}_p \simeq S^1$. The momentum in that case becomes an angular variable $0 \leq \theta \leq 2\pi$.

Going yet one step further, we can also ask what happens if the position space is *discrete and finite*, for example cyclic like the corners of a polygon, then interestingly enough the sample space of the analogue of a momentum observable associated with the particle hopping from one state to another would also becomes periodic and discrete. We have already run into the simplest example of this, a space with two points being just a classical bit, or classical Ising spin, which as we saw on a quantum level gives rise to the qubit or quantum spin. For that case it turned out that the position observable Z had two eigenvalues ± 1, and the same was true for the 'momentum operator' X. As you see, we have managed to wrap a whole lot of quantessence in a qubit, and will continue to do so. This concludes our discussion of a first crucial difference between the classical and quantum sampling spaces of a 'particle'.

Incompatible observables. The second difference is far more dramatic. It turns out that the position and momentum observables are incompatible, which means that a consistent framework for the quantum particle can only be based on either the momentum observable *or* on the position observable.[6] So, in going from classical phase space to the quantum space one can chose the momentum sample space indicated in Figure II.2.15 or for example the position space of Figure II.2.16, and we 'loose' the orthogonal dimension. The amputation of half the number of dimensions is quite an operation and I can imagine that you, following our discourse, may suffer from a kind of 'phantom pain' like experience. This loss implies a quantessential restriction on what can be considered 'a meaningful statement' about properties of the system, and at the same time creates ample room for void statements and 'fake news.'

What we just said also means that the quantum extension of our deductive logic gets severely restrained. Clearly if we compare the possible properties of a classical particle illustrated in Figure II.2.14 to the possible properties of a quantum particle given in Figure II.2.16, these are radically different. *Most importantly we cannot assign properties to the* P *and* X *observables simultaneously*, and hence cannot carry over the classical picture at all. What is left on the quantum level is that we may assign properties and ask for their conjunctions as long as they refer to one of the two observables, and this is illustrated in Figure II.2.16 where we did define two propositions A and B pertaining to the position variable and their logical conjunctions $\mathrm{A} \wedge \mathrm{B}$ and $\mathrm{A} \vee \mathrm{B}$. In conclusion, we note once more that because quantum operators in general do not commute, axes prominently present in the classical picture may be completely absent on the quantum level. This does not mean that the 'lost' observable X or P has taken the value zero and we have left out the corresponding axes. No, it says that a variable which is not part of the framework has no meaning let alone a value, and the axis is just not there!

We will run into these kind of situations repeatedly, where before making any strong statements on the properties of a state of a quantum system, we have to be explicit about the framework we are using. In quantum theory we apparently have one complete, consistent and rigorous mathematical formalism that supports many logically distinct frameworks. This may remind you of special relativity where one also distinguishes many reference frames which are relativistically equivalent, as they can be transformed into each other by a Lorentz transformation. But to make an argument you better do not mix up statements that hold in different frames. And here we are finding many frameworks which are quantum (or unitarily) equivalent but making a physical argument, you better stick to one if you want to keep your physics straight.

This may at first sight look strange and unfamiliar and a heavy load of reader unfriendly jargon, but at the same time it is a precise, concise and explicit statement of what states, dynamical variables and measurements in quantum theory are about. And it is this core structure of the theory that we want to extensively explore in the remainder of this volume. This exposition has hopefully made you feel more comfortable with it, because from the underlying mathematical structure lots of quantessential properties can be derived. These quantessential properties, which to the classical mind may appear exotic to say the least, are falsifiable at least in principle, and have turned quantum physics into a full-fledged scientific theory. The construction of this solid mathematical framework was largely the brilliant work of the second generation of outstanding quantum physicists, like Werner Heisenberg, Erwin Schrödinger, Paul Dirac, Max Born and John von Neumann to mention a few.

[6] In fact one may choose any linear combination of the two, but for the moment we choose this simple restriction.

The case of a quantum bit

Philosophers talk about an *ontology* in which the quantum reality could be understood and categorized. What are its basic entities, what are their measurable properties and what are the rules governing them? One likes to understand what the propositions or properties are that are either true or false. And as we have seen in quantum theory the rules about observables appear to be rather bizarre, and therefore it is illuminating to study their logical structure in more detail.

Projection operators. It is convenient to go back to some of the statements we made on page 292 of the previous section. Suppose we have some Hilbert space $\mathcal{H}$ and a suitable set of observables that are mutually commutative and their common eigenvectors $\{|i\rangle\}$ span $\mathcal{H}$. Or we could construct a single observable which would be non-degenerate and therefore satisfy

$$A|i\rangle = a_i|i\rangle\,,$$

with all its eigenvalues a_i being different. Then we could consider the *elementary projectors*:

$$P_i = |i\rangle\langle i|\,,$$

which satisfy:

$$\Sigma_i\ P_i = \mathbf{1}\,,$$

and therefore we can introduce its *logical negation* $\neg P_i = \mathbf{1} - \Sigma_{j\neq i}\ P_j$, which is of course also a projection operator that projects states on the subspace orthogonal to $|i\rangle$. These projectors all commute; furthermore the observable A can in this basis simply be expressed as

$$A = \Sigma_i a_i|i\rangle\langle i|\,,$$

with the eigenvalues as coefficients. The Hamiltonian operator for example can be written as:

$$H = \Sigma_n E_n\,|\psi_n\rangle\langle\psi_n|\,. \qquad \text{(II.2.18)}$$

Let us verify some of the equations above for the Pauli matrices. The projection operators would correspond to the matrices:

$$P_1 = |1\rangle\langle 1| = \begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix},\ P_{-1} = |-1\rangle\langle -1| = \begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix}. \qquad \text{(II.2.19)}$$

These operators commute and indeed $P_1 + P_{-1} = \mathbf{1}$. The observable Z can be expanded in the projection operators as $Z = P_1 - P_{-1}$. Just for completeness we also give the expressions related to the observable X:

$$P_+ = |+\rangle\langle +| = \frac{1}{2}\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix},\ P_- = |-\rangle\langle -| = \frac{1}{2}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix},$$

and similar properties hold.

With these projectors we may now associate properties or propositions that may be true or false in the sense that if we measure A and obtain some particular outcome a_k, stipulating that P_k is 1 (true), and all other P_i are 0 (false):

$$\begin{aligned} P_k|k\rangle &= 1\,|k\rangle \\ \neg P_k|k\,\rangle &= (\mathbf{1} - \Sigma_{j\neq k}P_j)\,|k\rangle = 0\,. \end{aligned}$$

You may verify this outcome from the examples above.

Non-commuting projectors. So far so good, but what happens if we want to define elementary conjunctions between properties, say we want to ask whether P *or* Q $(P \vee Q)$ is true. From Table II.2.1 one learns that such a proposition would correspond to the truth value of the projector PQ or QP. The logical proposition P *and* Q, $(P\wedge Q)$ has truth value $P+Q-PQ$, and also involves the product. But now we run into a problem because the product of two projectors is again a projector only if they commute. So in quantum mechanics neither PQ nor QP can in general be true or untrue, and this poses a fundamental problem from an ontological point of view.

Consider in the qubit example above, for instance the proposition $P_1 \vee P_+$. This would have to correspond to the prod-

uct operator

$$P_{1+} = P_1 P_+ = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \text{ or } P_{+1} = P_+ P_1 = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix},$$

but these are different and moreover neither of them is a projection operator ($P_{1+}^2 \neq P_{1+}$) to which truth values could be assigned. In the language used before we say that Z and X are indeed incompatible observables.

The choice of a framework. We can now avoid some of this by demanding that we only use a set of mutually commuting projectors or a set of compatible observables, linked to a given basis defined by some generic observable. Such a framework does indeed limit the number of properties that can be assigned to the system. But adopting such a framework one can use ordinary deductive logic concerning the restricted set of properties of the system.

And conversely a state can only have or not have a property a_i if we work in a framework where we can assign a truth value to its associated projector P_i. So, other non-commuting observables simply have no meaning in such a framework. And we have to think of such states in terms of a probability amplitude over the sample space connected to the framework one happens to be working with. There are many inequivalent such sets and it depends on what aspects of the theory one wants to study which one to choose. This observation suggests the use of the notion of a *single framework*, as a set in which to describe quantum states and also the propositions about the system which are meaningful in that framework. This defines an additional *syntactic rule* which forbids employing incompatible frameworks into a single description of the properties of the system. This is central to what is sometimes referred to as the *new quantum logic*.

In this single framework setting of quantum mechanics we return as closely as possible to a classical description of states with definite properties and statistical distributions over sample space. Describing the dynamics in such a single framework makes the quantum time evolution into some quite ordinary stochastic process as we will point out later.

Certain uncertainties

> Nothing [in quantum theory]... was more startling than Heisenberg's uncertainty principle, which denied the possibility of simultaneously measuring certain properties of motion. The uncertainty principle introduced us to quantum fluctuations, revealing empty space to be in fact a cauldron of activity.
>
> *John Archibald Wheeler, Geons, Black Holes & Quantum Foam* (1998)

Early on in the development of quantum theory it was Werner Heisenberg who proved his fundamental uncertainty relations *stating the impossibility of simultaneously measuring certain variables that characterize the state with arbitrary precision. There is a fundamental limit to the accuracy of quantum measurements set by Planck's constant. These relations, more than anything else, express the profound difference between classical and quantum systems. We discuss the position-momentum uncertainty relation for a particle state, and work out the detailed example for a qubit.*

Momentum versus position. Accepting that the state is completely specified by a wavefunction that will only tell you the probability amplitude for finding certain outcomes for any given observable another question remains: what does the wavefunction say about the momentum of the particle? There is no mention of momentum, it doesn't seem to play any role whatsoever in the definition of the state. This seems perfectly alright in view of what we have been talking about in the previous section on compatible observables and frameworks. All true, but could I not per-

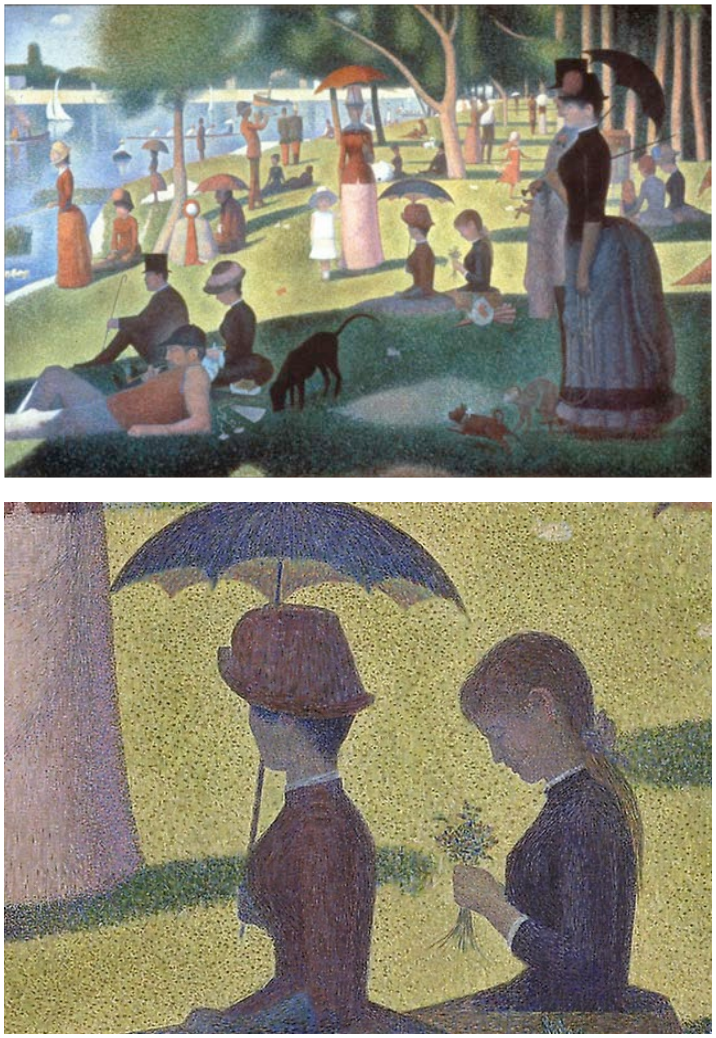


Figure II.2.17: *Pointillism.* Detail (bottom) of the pointillist painting *'A Sunday Afternoon on the Island of La Grande Jatte'* (top) by the French painter Georges Seurat. Painted some years before the moment when Planck made his groundbreaking quantum hypothesis, this work showed how a closer look may reveal a quantum structure.(Source: Wikimedia.)

fectly well decide to go out and just measure it, couldn't I? Yes, you certainly can and you would indeed get a definite answer. But the story is the same as with the position measurement. Say, if you prepare a particle in a certain state described by some wavefunction $\psi_0(x)$ and you measure a value for the momentum $p = p_0$. Then you could repeat the whole procedure and somehow again prepare the particle in exactly the same initial state and then once more measure its momentum, what would you find? Well, the statement is that in general you would get another outcome $p_1 \neq p_0$. How vague can a theory be? Well, in a sense that's precisely what quantum theory is about, it tells you exactly how vague outcomes of measurements are.

Certain uncertainties. Probabilities imply uncertainties in outcome, but the magnitude of those uncertainties are precisely determined. We have to deal with 'certain uncertainties' so to speak. In fact there are strong bounds on the uncertainties of different observable quantities. You might for example try to circumvent the quantum uncertainties by being smart. If you say, I measure the position of a particle so that it is well localized in position space, and then immediately after I measure the momentum so that I can also localize the particle in momentum space. By doing this, am I not arbitrarily close to the state in classical physics where we could assign a precise position and momentum to a particle at any instant? The stupefying answer is: certainly not!

The Heisenberg uncertainty principle

The quantessential message on the differences between classical and quantum observables is very clearly, concisely and quantitatively encoded in what are called the Heisenberg uncertainty relations. For the case at hand he derived that for any state of a particle the following relation holds for the uncertainties in position Δx and momentum Δp of the particle *in that state*:

$$\Delta x \, \Delta p \geq \frac{h}{4\pi} ,$$

where the spread is just the width of the respective probability distributions. It relates measurement outcomes for

the same state in different frameworks! What Heisenberg proved was exactly that there is a lower bound on the product of those widths. It shows unequivocally that the situation, generally assumed in classical physics, where both widths can be taken to zero in principle (assuming ideal measurement apparatus etc.) is not possible in quantum theory as a matter of principle.

If we drop a marble in a bowl, it will after some oscillations settle down in the minimal energy state which means that it will be at rest at the bottom of the bowl. Momentum zero and position fixed exactly: no uncertainties. Classically yes, but because of the uncertainty relations, or the particle-wave duality for that matter, this cannot be the quantum story. A quantum marble cannot settle down in a state where it is at rest at the bottom of the quantum bowl, because then its position and momentum would be exactly known, there would be no uncertainty, and that is not an allowed state. The lowest energy state of the quantum marble in a quantum bowl turns out to be one where the uncertainties in position and momentum are about equal and saturate the lower bound of the uncertainty relation. It gets as close to the classical ideal as possible you could say, but the truth is that the lowest energy state of the particle does not specify where it exactly is nor what its momentum precisely is.

As we will see later, there exist Heisenberg uncertainty relations between any pair of observables A and B, only if a non-trivial (non-zero) bound only occurs for an incompatible (non-commuting) pair. What does this have to do with my expose about frameworks? Surprisingly little in fact. The uncertainty relations link the variance in outcomes of measurements of a pair of observables in any given state. So given a state $|\psi\rangle$ of a particle, one can imagine making many independent measurements of say the position x of the particle in that state. This of course does not mean that you make a simple sequence of measurements on a single particle, because a measurement will *change*, what do I say, will *collapse* the state! So you have to prepare

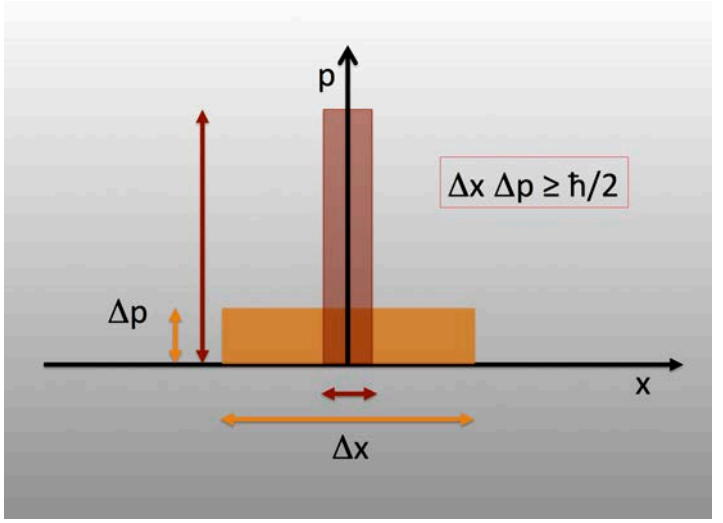


Figure II.2.18: *Heisenberg's uncertainty relation.* The uncertainty relations for position and momentum define the minimal area in classical phase space corresponding to possible states with uncertainties $\Delta\mathrm{x}$ and $\Delta\mathrm{p}$.

'identical' particles in identical states and then make repeated measurements of the observables in question. You may start with position to obtain an average or expectation value $\bar{\mathrm{x}}$ and some variance Δx. Subsequently, one could make independent momentum measurements producing a distribution of outcomes with an average $\bar{\mathrm{p}}$ and variance $\Delta\mathrm{p}$. Heisenberg's fundamental relation says that the product of these variances or 'uncertainties' is larger than or equal to $\hbar/2 = \mathrm{h}/4\pi$. So we do not compare individual measurement outcomes but distributions thereof. In Figure II.2.18 we show that the product of uncertainties in a given state corresponds to a certain rectangular area in the (classical) phase space, the shape of the rectangle depends on the state but its area has to be larger than the minimal area indicated in the figure. The conclusion therefore is that in the quantum world there can be no states in which both position and momentum take on precise values! It is a profound statement concerning probabilities of measurement outcomes of different variables in any given state, but that in itself has no bearing on the logical struc-

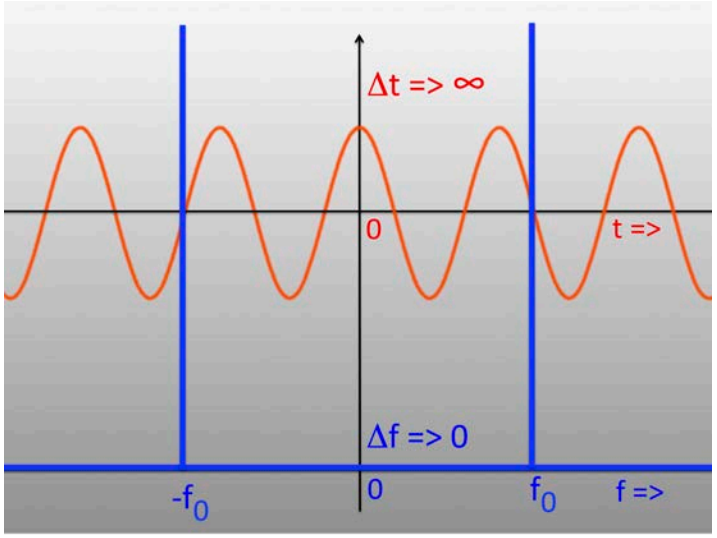


Figure II.2.19: *Time-frequency duality.* Representation of a sound signal as a periodic function of pressure (in red) in time: $P(t) = \cos \omega t$, or as a function of frequency $P(\omega)$ with two narrow peaks around $\omega/2\pi = \pm f_0$. A periodic signal is not localized in time with $\Delta t \to \infty$, but is very localized in frequency, $\Delta f \to 0$.

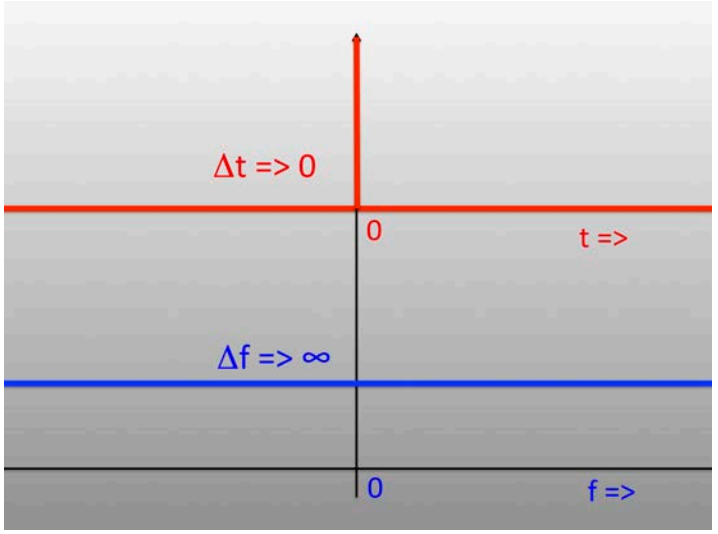


Figure II.2.20: *Time-frequency duality.* The 'clap in hands' signal is very much localized in time, $\Delta t \to 0$ and spread out very widely in the frequency domain, $\Delta f \to \infty$.

ture of the quantum world we were discussing in the previous section, though it is of course consistent with it.

Non-trivial uncertainty relations exist for all pairs of *incompatable* or non-commuting observables, because these cannot be measured simultaneously, or stated more precisely: if the system has a definite value for the one variable, it is not possible to assign a value for the other. One can choose either one to quantify or describe *any* state of the system but not both. We conclude that quantum states are thus described by a maximal number of mutually compatible observables that define a framework. And indeed not all choices of sets of compatible observables are equally convenient or practical, that depends on what you want to know about the system.

A sound analogy

In this subsection we take one further step trying to understand what incompatible observables, and the uncertainty relations they obey, mean. Surprisingly enough, there are uncertainty relation look-alikes in the classical physics of waves that may take some of the mystery away. Let us for example think about sound. Sound is a pressure wave that passes. At some point in space we hear a sound signal and ask how we would characterize it. One way is to plot the pressure variations in real time, and another way is to represent the signal in the frequency domain as a superposition of sounds of different frequencies with different amplitudes. These pictures would look quite different but contain the same information and are just different representations of the same signal.

Let us first look at (or listen to) a pure tone like the 'a'. A truly pure 'a' of 441 hertz is represented in time by a pure sine or cosine wave of a fixed wavelength which has

that single frequency of 441 Hz. But for a cosine to be pure it has to last a very, very long time (compared to the inverse frequency), as I indicated with the red curve in Figure II.2.19. So, a pure tone is very much extended in the time domain, but if you look in the frequency domain it is extremely narrow because the signal has only a single frequency (in fact $f = \pm f_0$) as you see in the narrow peaked blue curve in the same figure. Now, in Figure II.2.20, the opposite happens when I clap loudly my hands once, or shoot a gun, then the signal is extremely short in the time domain, but in the frequency domain it is very wide.[7] If I clap my hands or bang a hammer on the table and I ask you what the pitch was of the sound you heard, you will answer that you could not determine any pitch because the sound lasted for too short a time. If you were to fire a revolver next to a piano and keep the right pedal down then all the strings will resonate showing that basically all the frequencies were present in the sound of the shot: an overdose of pitch rather than no pitch. The upshot of this exercise is that indeed duration and frequency are dual to each other. The more accurate the frequency (i.e. the smaller Δf) in a signal, the longer it has to last (i.e. the larger Δt) and vice versa. In other words one expects a relation like $\Delta f \Delta t \geq constant$ to hold. This is true and by the way the constant is $1/4\pi$. The lesson here is that you can't have it all: you cannot have the cake and eat it. The physics in this example is quite comprehensible and much what we experience in daily life, yet we encounter a situation where we cannot ask for a signal that is precisely localized in time and also has a well-defined pitch. These two physical quantities are in that sense incompatible, and this duality is intimately linked to the wave character of the phenomenon.

Let us switch now to electromagnetic waves which are made up of many photons. Remember that photons obey the Planck-Einstein relation $E = h\nu$, so we can replace the frequency ν by the energy and obtain an energy-time relation $\Delta E \Delta t \geq \hbar/2$, and that is indeed exactly an instance of Heisenberg's uncertainty relations. The interpretation is that we cannot measure both variables with arbitrary precision simultaneously.

[7] The two figures are not entirely symmetric because I choose to clap at time $t = 0$, the exactly dual situation would be obtained by choosing $\omega = 0$ in the first figure then the cosine function would become constant, $\cos 0 = 1$, and the two peaks move on top of each other as $f_0 = 0$.

Heisenberg's derivation

With the formal ingredients we have so far introduced it turns out to be rather straightforward to actually derive the uncertainty relation for two observables. It really is a matter of simple algebra but with objects that look awesome. You feel like you are juggling with antique Chinese vases but in fact they are just empty plastic bottles.

Let us consider two observables A and B, in particular we study two vectors $(A-a)|\psi\rangle$ and $(B-b)|\psi\rangle$ where $a = \langle A\rangle$ and $b = \langle B\rangle$ are real numbers. The variance (the mean square deviation) of an operator A in a state $|\psi\rangle$ is defined in terms of expectation values as (see the *Math Excursion* on Probability and statistics in Volume III):

$$(\Delta A)^2 \equiv < (A-a)^2 > = < A^2 - 2aA + a^2 > = < A^2 > - a^2 .$$

The variance is a measure for the width of the distribution. Note that if $|\psi\rangle$ is an eigenstate of A, meaning that $A|\psi\rangle = a|\psi\rangle$, then $\Delta A = 0$. Now there is a famous inequality for vectors called the Schwarz inequality. It says that if you have two vectors and their inner product, then the product of their lengths squared is always larger or equal than their inner product squared. In the familiar Euclidean setting we would have $|\mathbf{v}\cdot\mathbf{w}|^2 = |\mathbf{v}|^2|\mathbf{w}|^2\cos^2\theta \leqslant |\mathbf{v}|^2|\mathbf{w}|^2$, which holds because the cosine squared is smaller than one. Applied to our vectors above this yields the statement that

$$\langle |A-a|^2\rangle \, \langle |B-b|^2\rangle \geq |\langle (A-a)(B-b)\rangle|^2 .$$

Note that on the right-hand side $\langle (A-a)(B-b)\rangle$ is just some complex number, let us call this number z. Then the absolute value squared is

$$|z|^2 = z^* z = (\mathcal{R}e\, z)^2 + (\mathcal{I}m\, z)^2 \,,$$

and clearly $|z|^2 \geq (\mathcal{I}m\, z)^2$, where

$$(\mathcal{I}m\, z) = \frac{1}{2i}(z - z^*) = \frac{1}{2i}\,\langle [A,B]\rangle \,.$$

The commutator is the only term that survives because $z^* = \langle (A-a)(B-b)\rangle^* = \langle (B-b)(A-a)\rangle$ and all other terms cancel out.

Putting the results of the above equations together, we arrive at the desired result, the celebrated *Heisenberg's uncertainty relation* in its general form:

$$\Delta A\ \Delta B \geq \frac{1}{2}|<i[A,B]>| \,. \qquad \text{(II.2.20)}$$

Note that if A and B are hermitian then also $i[A,B]$ is, which makes its expectation value real. We obtain a non-zero lower bound for the product of uncertainties in the case the operators A and B do not commute. An immediate consequence of the relation is that in any state the uncertainty in the measurement value for two such incompatible variables can never be zero for both. There is a complementarity: the more precise you know observable A the less precise you know the value B. It is the golden rule for giving and taking: you can't have it all. ■

Qubit uncertainties

After this derivation of the precise form (II.2.20) of the uncertainty relations it is interesting to see how these relations play out for the simple case of qubits.

We are going to check the qubit uncertainties in the cases we considered before. If we take as two incompatible observables $A = Z$ and $B = X$, then the relation would read

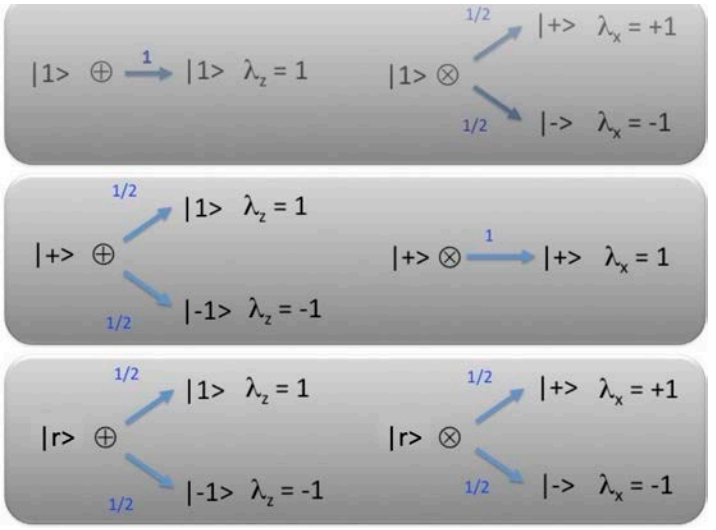


Figure II.2.21: *Spin uncertainties.* Uncertainty in spin measurements of Z and X denoted by $\oplus$ and $\otimes$ respectively, for the states $|1\rangle$ and $|+\rangle$ respectively. The blue numbers are the probabilities for the various outcomes. We see that where one of the spin measurements has minimal uncertainty ($\Delta = 0$), the other is maximal ($\Delta = 1$). Had we chosen an eigenstate $|r\rangle$ of Y then the uncertainty in both X and Z would have been maximal, and the uncertainty relation would again be satisfied.

$$\Delta Z\ \Delta X \geq \frac{1}{2}\,|\langle\ i[Z,X]\ \rangle| = |\langle\psi|Y|\psi\rangle| \,. \qquad \text{(II.2.21)}$$

Let us then choose for the states $|\psi\rangle$ subsequently (i) $|1\rangle$, (ii) $|+\rangle$, and (iii) the eigenstate of Y with eigenvalue $+1$, denoted by $|r\rangle$. We recall that $Z^2 = X^2 = 1$ and also that $|\langle\ A\ \rangle|$ equals either 1 or 0 for our A depending on whether $|\psi\rangle$ is an eigenstate of A or not. This makes the calculation relatively simple for example for the left-hand side we obtain:

$$\begin{aligned}(\Delta A)^2 &= \langle A^2\rangle - (\langle A\rangle)^2 \\ &= \begin{cases} 1-1^2 = 0 \text{ (if eigenstate)} \\ 1-0^2 = 1 \text{ (if not eigenstate)} \end{cases} ,\end{aligned}$$

and for the right-hand side:

$$|\langle\psi|\ Y\ |\psi\rangle| = \begin{cases} 1 \text{ (if eigenstate)} \\ 0 \text{ (if not eigenstate)} \end{cases} .$$

So, for the subsequent cases we end up with the following inequalities (i) $0 \cdot 1 \geq 0$, case (ii) $1 \cdot 0 \geq 0$ and case (iii) $1 \cdot 1 \geq 1$, and we happily agree that in all cases the uncertainty relation is satisfied and moreover saturates the lower bound. In Figure II.2.21 we give the various measurement outcomes with their probabilities for the Z and X observables for the three states $|1\rangle$, $|+\rangle$ and $|r\rangle$.

Ground state energy. For a quantum particle the lowest energy state will, even if it is weakly localized, always have some extra *zero point energy* associated with it. Adding up all the zero point energies of all particles means that what we call the 'vacuum' must be full of energy. Can't we get it out and do something useful with it is a question that regularly comes up. No presumably not. All physical observables like spectral lines and so on are related with energy differences, and you are free to choose the ground state level as it has no observable effect.
Having said that, you could of course scratch your head, and modestly point out to me that there is a notable exception, and that is Einstein's theory of general relativity, where the vacuum energy does indeed cause physical effects, even of cosmic importance. The shocking news has been that indeed the energy balance in our universe is dominated by the vacuum contribution, which amounts to some 70 percent. But it remains a complete mystery why that number is what it is. Yet, this vacuum energy is like a cosmological constant and it has a mind-blowing property that it anti-gravitates and exerts an outward gravitational pressure that makes the universe expand, and will keep the universe expanding forever as we discussed briefly in Chapter I.2. So, there are instances that much ado about nothing is quite OK, especially if one understands nothing about that nothing. □

From these the variances on the left-hand side of (II.2.21) can immediately be read.

Let me make a final comment. Let us go back to the discussion of 'bit dynamics' at the beginning of Chapter II.1. There we stated that Z could be interpreted as a 'position' operator giving the ± 1 eigenvalue for the spin-up (down) state. In that context the X operator 'generated' translations (hopping in z) and as such acted like a 'momentum' operator. And once more we see that the two operators do not commute and hence satisfy non-trivial uncertainty relations. By the way, these uncertainties imply that quantum computers will provide an array of potential answers, from which the correct one has to be selected somehow.

The breakdown of classical determinism

The uncertainty relations imply strict limits on the predictability in physics. This unpredictability implies the breakdown of classical determinism. A surprising and profound philosophical sacrifice in the realm of our material universe.

The uncertainty relations of quantum theory go further: they imply that if we know the particle has a small uncertainty in position because we just measured its position, then it is in a state where the uncertainty in momentum will be relatively large. If you were to ask me to tell you where the particle would be some time after, then it would be hard to point at a specific point. I do know its starting position precisely, but I don't know its momentum, and thus it is hard to say where it goes and with what momentum. We see that the quantum postulates, concisely expressed in the uncertainty relations, imply the breakdown of classical predictability and determinism. This is one more truly quantessential feature of the underlying reality.

Humankind's limited abilities to observe have through our

common experiences precipitated in what we call deep intuitions about how the world works. And such intuitions tend to shape our judgements and expectations. One thing that has become inescapably clear is that quantum theory has shown such intuitions to be essentially mistaken in an essential way, a sobering thought indeed. That one more illustrates the power of the invisible. At this point I should remind you of the wonderful quote from the Feynman's which I included in the preface to Volume I on page xiv.

This fundamental indeterminacy in nature has lead to numerous speculations on the far-reaching consequences it might have, varying from metaphysical hocus-pocus like floating tables to explanations of the human free will.

Why does classical physics exist anyway?

After all this classical physics bashing, you might ask: how come classical physics is doing so extremely well in ordinary life, if it is so fundamentally wrong? How can that be?

A golf ball. Let us consider a golf ball. If I neglect its internal structure, should I not treat it as a quantum particle and if I do so just reproduce the classical answer? Yes, you better do so, otherwise quantum theory would be in conflict with direct observations. Suppose you would make an extremely accurate measurement and measure its momentum in four decimal places so $\Delta p = 10^{-4}\ \mathrm{kg\,m\,s^{-1}}$, then substituting this into the uncertainty relation you would find that the uncertainty in position would be a mesmerizingly tiny $\Delta x \geq \hbar/4\pi\Delta p \simeq .5 \times 10^{-30}\mathrm{m}$. But wait, that is the realm where string theorists wander. You will agree that nobody is ever going to make a measurement of position with such 30-decimal places accuracy, let alone of a golf ball! Think of an ultimate machine like the Large Hadron Collider at CERN, where physicists are able to localize particles 'only' up to about 10^{-18} meters at present. Physicists may have their ways, but to verify the uncertainty relations by playing golf in the LHC is not of them. So, what then saves the day for classical physics or if you prefer, what saves quantum physics? That is the dazzling smallness of Planck's constant if you express it in our anthropocentric system of units, made up of meters, seconds and kilograms. That is why the basic need for quantum theory, i.e. the failure of classical theory manifests itself at first only on small scales, and it is also for that reason that it took so long for the quantum world to be discovered.

An electron. To appreciate the point just made, let us replace the golf ball by an electron with a mass of about $10^{-30}\ \mathrm{kg}$. Then we could easily measure its momentum with an uncertainty of $10^{-30}\ \mathrm{kg\ m\ s^{-1}}$, leaving a position uncertainty of about one tenth of a millimeter. So, indeed in an atom with a typical size of $10^{-10}\ \mathrm{m}$ – one-tenth of a nanometer – this uncertainty matters and therefore we should treat the electron quantum mechanically. This observation by the way implies that we should no longer think of electrons as well-localized particles orbiting the nucleus. Indeed the way the atom is usually depicted (see Figure I.3.6) is a severe misrepresentation inherited from our classical intuition. Rather we should represent the electron as a standing wave pattern of the probability wave in the tiny volume of atomic size. Atoms are not like tiny solar systems, but rather like tiny *quantum bongos!*
In fact knowing the size of the atom to be about $\Delta x \sim 10^{-10}\ \mathrm{m}$ one may use the uncertainty relations to estimate the minimal momentum as $p \sim \Delta p = \hbar/(4\pi\Delta x) \simeq 10^{24}\mathrm{kgm/s}$, which corresponds to an electron energy of $10^{-19}\mathrm{Joule} \simeq 1\ eV$. And 1 electron Volt is indeed the order of magnitude of atomic energy levels. It can't be much less and you could even say that this is one of the reasons that matter is actually stable.

The emergence of classical physics. The macroscopic world which obeys by definition the classical laws of physics is a world consisting of emergent phenomena, and the

classical laws are therefore only approximately true. The world we perceive is an incredibly coarse-grained version of a well-shielded microscopic reality. Our world has an incredible amount of entropy exactly because there is so much information hidden within, and science is exactly the systematic uncovering of that information and making it accessible. It is a gigantic hacking operation, a gigantic striptease of mother nature in which she slowly confides to us her deepest secrets. There are many *why*-questions one may ask on the macroscopic level that can be answered only after they have been turned into *how*-questions on the underlying quantum level. In other words, classical physics is the emergent macroscopic manifestation of an underlying quantum world. The *quantessence* comprises of the unescapable laws underlying classical reality. This exemplifies the profound gain of progressing insight in the long run. The process of scientific progress is seldomly gradual and smooth, and rather proceeds unpredictably, with sudden shocks. In evolutionary biology Jay Gould introduced the notion of *punctuated equilibrium*, which clearly echoes in the picture of long periods of 'normal' science, broken up by scientific revolutions, radical turning points in our thinking leading to paradigm shifts, as described by Thomas Kuhn in his book on *Scientific Revolutions*. I may add that important novel cultural dimensions have opened up, as a result of this process of progressing insight in science as I have argued in my book *In praise of science*.

Further reading on quantum measurement:

- *Quantum Theory*
  D. Bohm
  Dover Publications Inc (1989)
- *Quantum Measurement Theory and its Applications*
  Kurt Jacobs
  Cambridge University Press (2017)

Table II.2.2: **Key quantum principles introduced in this chapter on observables.**

	Keyword	*Description*
(ii)	**Observables**	A physical variable a or observable is represented by a hermitian operator or matrix A. To the system as a whole corresponds a set (algebra) of observables $\mathcal{O} = \{A, B, \ldots\}$.
(iii)	**Eigenvalues**	The observable A has a set of real eigenvalues $\{a_i\}$ which make up the sample space or spectrum S_a of possible measurement outcomes for A.
(iv)	**Eigenvectors**	To each eigenvalue a_i corresponds an eigenvector $\|a_i\rangle$, or a subspace $\mathcal{V}_i^a$.
(v)	**Preferred frames**	In the non-degenerate case, the eigenvalues of A are all different, their number equals the dimension of the Hilbert space, and the set of normalized eigenvectors $\{\|a_i\rangle\}$ forms an orthonormal basis for $\mathcal{H}$.
(vi)	**Superposition**	Any state $\|\psi\rangle$ has a linear expansion in the basis of any framework. $\|\psi\rangle = \sum_i \beta_i \|a_i\rangle$.
(vii)	**(In)compatibility**	Observables are compatible if (and only if) they mutually commute so that common eigenvectors can be chosen. Observables that do not commute are by definition incompatible.
(viii)	**Frameworks**	A maximal number of independent compatible observables forms a framework $\mathcal{F}$. A complete orthonormal set of joint eigenvectors of a framework forms a basis for the Hilbert space $\mathcal{H}$.
(ix)	**Measurement outcomes**	When making a measurement of an observable A on a state $\|\psi\rangle$ there is a probability $p_k = \|\langle a_k \|\psi\rangle\|^2$ of obtaining the result a_k.
(x)	**Projective measurement**	Upon measuring the value a_i in a strong or projective measurement of A, the state $\|\psi\rangle$ 'collapses' to the eigenstate $\|a_i\rangle$ of A. This statement is referred to as the *projection postulate* of Von Neumann.

Chapter II.3

Interference

We have seen that a quantum particle like an electron has wave-like features and that an electromagnetic wave has particle-like properties as we may consider such a wave as a collective of photons. This naturally raises the question how quantum particles really exhibit these wave-like properties. In this chapter we focus on the question of whether particles can show interference effects like waves do. The answer to this question is affirmative, as is demonstrated by the famous double slit experiments of various kinds. In this chapter we consider classical as well as quantum wave phenomena.

Figure II.3.1: *Dew drop.* In this lithograph of M.C. Escher, the reflection of light causes the image of the windows of the observer's room. The refraction or breaking of light at water-air interface yields the enlarged image of the underlying veins of the leaf. (© 2023 The M.C. Escher Company.)

Classical wave theory and optics

Classical geometric optics treats light as straight rays that can be deflected or reflected by different media. The strict geometrical picture consisting of straight light rays can be augmented by the wave-type constructions based on Huygens' principle, which states that any point on a wavefront can be considered as a source of secondary spherical waves. It is not only the laws and patterns of geometric optics like reflection and refraction (breaking) of light at interfaces between different media that can then be explained, but also more subtle effects like diffraction (bending).

Basics of wave theory

Characteristics of waves. Let us recall some basics of classical wave theory. A propagating wave can in general be characterized by:
(i) a periodic wave pattern of subsequent *maxima and minima*. The height of the maxima is called the *amplitude* of the wave. The curves connecting adjacent maxima are

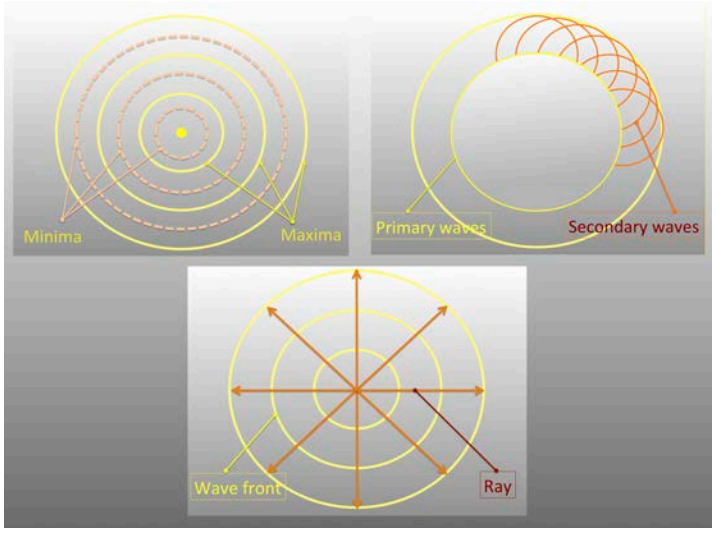


Figure II.3.2: *Wave patterns and propagation.* We show the wave pattern corresponding to propagating wavefronts from a single point-like source, to illustrate the some basic wave concepts.

called *wavefronts*; for the case of plane waves these are parallel straight lines or planes, while for a single source these are circular or spherical as illustrated in Figure II.3.2.
(ii) a pattern of *rays*, which are lines perpendicular to the wave fronts. So from a point source the rays are straight lines pointing radially out.
(iii) a *wavelength* λ, which is defined as the distance between two subsequent wavefronts measured along a ray. Often one uses the *wavenumber* k defined by $\mathrm{k} = 2\pi/\lambda$, instead of the wavelength.
(iv) a *speed* v, which is the speed at which the wavefronts propagate. For light and other electromagnetic forms of radiation propagating in vacuum, this is the universal speed of light c. In physical media (like glass) with electrodynamic properties different from the vacuum, however, the velocity of light will be less than its universal value in vacuum. The speed of light in media may generally depend on the wavelength (or frequency).
(v) a *frequency* f refers to the frequency by which every point in the wave oscillates.
(vi) we distinguish *longitudinal* and *transversal* waves where the medium oscillates parallel to the direction of propagation (sound), or orthogonal (light).
(vii) a *polarization*. Transversal waves can be (linearly) *polarized*, meaning that there is a single orthogonal axis along which the field oscillates.

Typical sizes and scales. For water waves the wavelength may vary from micrometers to many miles. For sound audible by the ear, in air at room temperature, the frequency f varies from $20\,\mathrm{Hz}$ to $20.000\,\mathrm{Hz}$; and with the sound velocity $\mathrm{v} = 343\,\mathrm{m/s}$, the wavelength would vary between $1.7\,\mathrm{cm}$ and $17\,\mathrm{m}$. For visible light the typical wavelength is thousands of angströms ($\sim 10^{-7}\,\mathrm{m}$). It is easier to remember for microwaves, because the wavelength you correctly guess to be of the order of micrometers. For quantum particle waves the scale is set by the De Broglie wavelength $\lambda = \hbar/\mathrm{p}$, typically about 10^{-10} meters or 1 angström.

Fundamental wave relations. There is a fundamental relation between the velocity, frequency and wavelength of a wave given by $\mathrm{v} = \lambda\,\mathrm{f}$. Mostly when talking about waves one assumes these are described by a linear theory. In such situations the linear superposition principle holds, so to understand the wave phenomena caused by independent sources one can simply add the wave patterns produced by the sources individually. On the one hand this applies in general to wave phenomena, as long as the oscillations are small because then the linear approximation holds well, but on the other hand we know that the Maxwell equations describing the electromagnetic waves are linear, and so are the Schrödinger and Dirac equations.

The physics of waves in a medium is interesting, because a wave carries a certain amount of energy and momentum. This, however, does not imply that matter somehow moves along with the wave. Think of a wave of water, you drop a stone in the pond which excites the water surface, locally perturbing the equilibrium situation. It is the de-

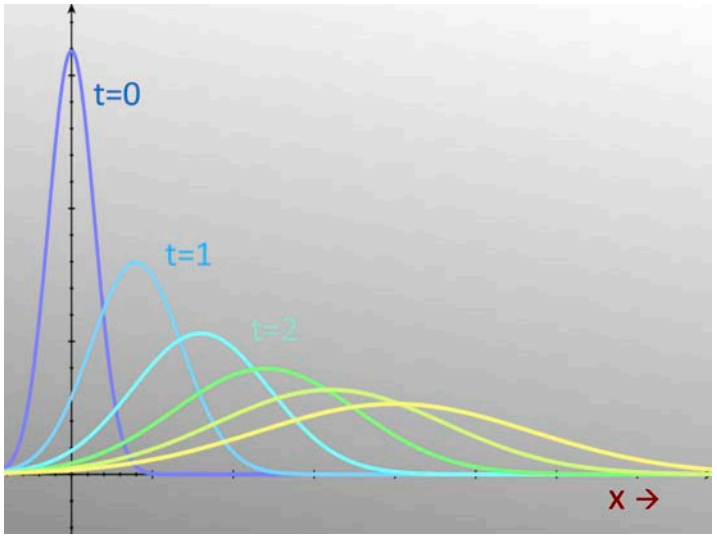


Figure II.3.3: *Dispersion of a wavepacket.* Depicted is a Gaussian shaped wavepacket in $\mathrm{x}-$space at $\mathrm{t} = 0$ and after five equal time intervals. The packet disperses (broadens) in time because different components travel at a different speed.

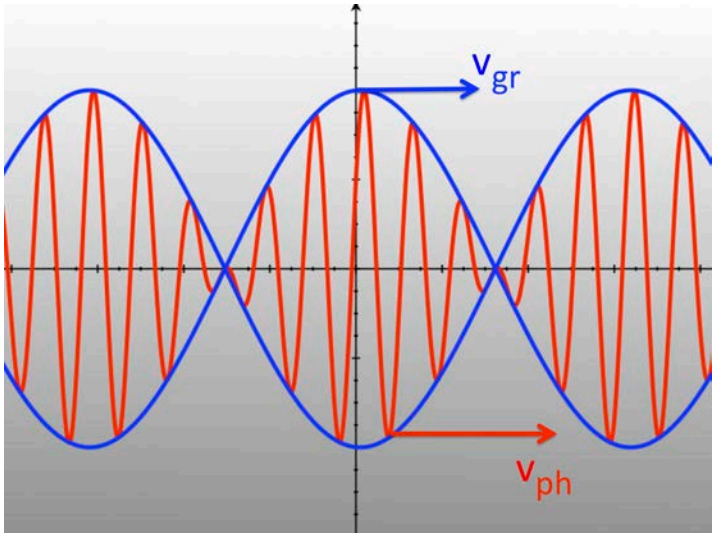


Figure II.3.4: *Group and phase velocity.* We have depicted a superposition of two linear waves with different frequencies and momenta. The combined result shows an *enveloping wave* in blue moving with the group velocity $\mathrm{v_{gr}}$, and the actual superposition in red moving faster with phase velocity $\mathrm{v_{ph}}$.

formation energy of the (elastic) medium that causes the perturbation (along with some characteristic deformation energy density) to spread as a wave pattern. As the total energy of the perturbation is conserved (if we assume that there is no dissipation), the amplitude of the circular wave has to decrease in time because the circumference of the wavefront increases. Anyway, for this transversal wave the position of a water molecule stays fixed as it only oscillates up and down. In the case of sound, the air molecules swing forth and back, but also in that case there is no material streaming along with the wave.

With light waves the situation is different though, because the lightwave is made up of photons, all moving with the same speed of light. The classical wave does not correspond to a single photon, rather it is a strange coherent superposition of different states with a different number of photons in them. They may all have the same frequency, but the various terms can involve quite arbitrary phases. As a matter of fact what this means is that the number of photons corresponding to a 'classical' wave is really not defined. This is not meant in a statistical sense but in a more fundamental way. To speak in the spirit of the previous chapter, their number is not defined, or better indefinite, because the corresponding 'number operator' is incompatible (does not commute) with the quantum operator that creates the classical wave configuration from the vacuum. In other words an electromagnetic wave is not in an eigenstate of the photon-number operator.

Dispersion. We have mentioned the fact that waves of different wavelength or frequency may travel at different speeds: this phenomenon is called *dispersion*. The most well known is the dispersion of light in glass for example, giving rise to separation of colors when light passes through a prism, as in Figure II.3.8. Dispersion means that the velocity and frequency depend on the wavelength or the wavenumber. It is usually specified by giving the functional relation between the angular frequency $\omega \equiv 2\pi\mathrm{f}$

and the wavenumber k, so by specifying $\omega = \omega(k)$. And we have seen that electromagnetic waves satisfy the linear dispersion relation $\omega = ck$, while for the De Broglie matter waves we have a quadratic dispersion because $E = p^2/2m$ with $p = \hbar k$ and $E = \hbar\omega$ yields: $\omega = \hbar k^2/2m$.

Broadening. The effect of dispersion manifests itself if we consider the time evolution of a *wavepacket*, which is just some linear superposition of components with different wavelengths. In Figure II.3.3 we see an initial packet that has some shape which is spatially localized with a certain width. One will find that such a packet will broaden or spread out (disperse) during its propagation, because the momentum components that make up the packet move at different speeds.

Group velocity. The next question that comes to mind is what the velocity of this wave packet is. After all it is made up of different components that move with different velocities. The basic answer to this question is illustrated in Figure II.3.4, for the simple case where we have shown the linear superposition of two waves with different frequencies and wave numbers, the combination can be rewritten as a product of a difference and sum wave with frequencies $\omega_\pm = (\omega_1 \pm \omega_2)/2$ and wave numbers $k_\pm = (k_1 \pm k_2)/2$. What we obtain is that the actual superposition, which is the wave pattern in red, propagates 'inside' the slowly moving enveloping wave in blue. You could say that the red wave with frequency ω_+ and wavenumber k_+ has a frequency modulated by the blue wave with ω_- and wavenumber k_-. The red wave moves with the phase velocity $v_{ph} = \omega_+/k_+$, whereas the envelope moves with the group velocity $v_{gr} = \omega_-/k_-$.

Dissipation. Dissipation refers to the loss of energy of a system, for example to the environment, or by producing heat internally due to friction. For waves, dissipation is often caused by inelasticity (viscosity) of the medium. Dissipation causes the signal to die out. Note that dispersion is not a dissipative phenomenon; it just is a consequence of the fact that different components of the wave packet move with different velocities.

Figure II.3.5: *Three views.* This picture offers three perspectives on yours truly, from a *direct, to the point*, a *reflective* and a *refractive* point of view. This can be achieved by just looking at a glass of wine!

Reflection, transmission, breaking and diffraction

Huygens' principle. To find out how the wavefront of a propagating wave moves forward, one may consider every point on the front as a source from which secondary waves emanate. The envelope of the secondary wavefronts defines the new wavefront. This is illustrated in the top right picture of Figure II.3.2. Huygens' principle is a powerful tool to explain all kinds of generic wave phenomena, like reflection, refraction, diffraction and interference. A nice example of reflection and refraction on which the working of lenses is based is provided by M.C. Escher's lithograph *The dew droplet* in Figure II.3.1.

Reflection. Light can be reflected off a surface, like in the reflection of an ordinary mirror. The law of reflection in

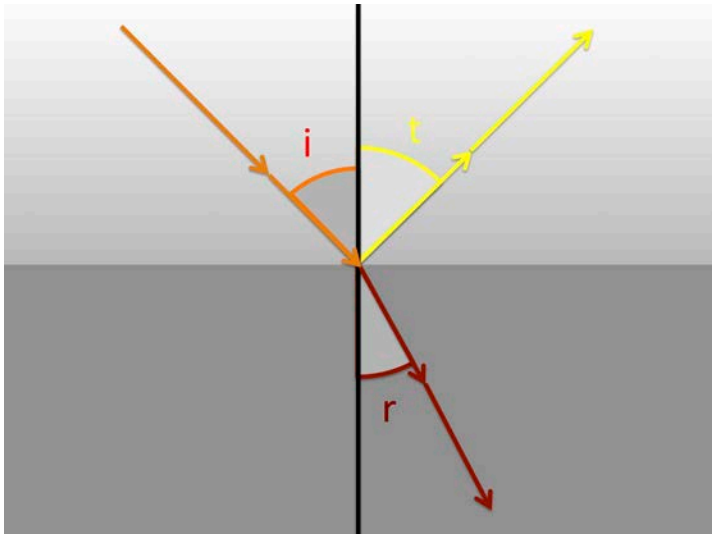


Figure II.3.6: *Reflection and refraction.* The picture illustrates reflection and refraction at an interface between two media. From the vacuum to any medium the angle with the normal to the interface of the refracted ray r is smaller than the angle i of the incoming beam.

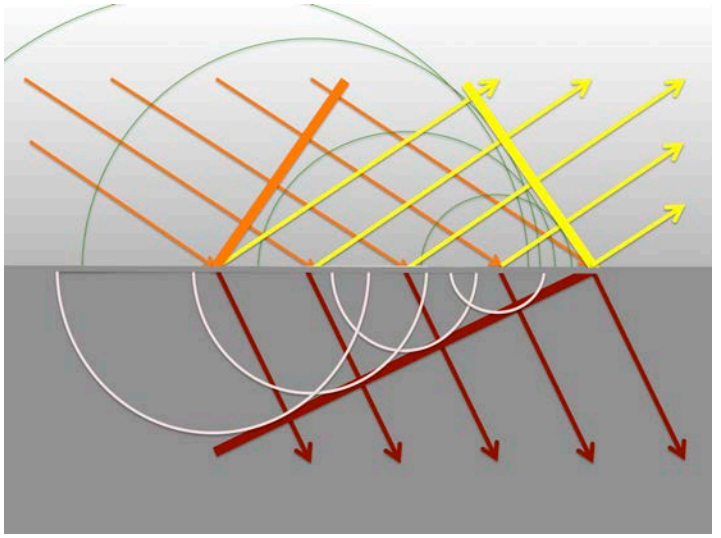


Figure II.3.7: *Huygens' principle.* The construction for the reflected and refracted beams (thin lines) and wavefronts (fat lines) using Huygens' principle, assuming you know the ratio of velocities in the two media, or breaking index.

geometric optics reads simply:

$$i = t\,,$$

or in words the angle i of the incoming beam (with the normal on the surface) is equal the angle t of the reflected beam. This is illustrated in fig II.3.6.

Refraction or breaking. The law for breaking of light at an interface between two media with relative breaking index n is given by Snellius' law which is also illustrated in the same figure:

$$\frac{\sin i}{\sin r} = n\,,$$

where n is given by the ratio of the speed of lights in medium 1 and medium 2:

$$n = \frac{c_1(f)}{c_2(f)}\,.$$

The proof of both laws can be given using Huygens' principle as we depicted in Figure II.3.7. We use the principle at the points where the incoming rays hit the layer between the two media, where the new front can be constructed using the same radii in the same medium (reflection), or reduced radii (because of the reduced speed of light) in the dense medium.

Note that whereas in vacuum the velocity of light is universal and therefore does not depend on the frequency or wavelength (color), this is no longer true in other media. As a consequence the angle of refraction will be different for different colors, as was so beautifully demonstrated by Newton by letting a sun ray pass through a prism (see Figure II.3.8).

Bragg diffraction and reflection.

William Henry Bragg and his son Lawrence Bragg proposed in 1913 a nice explanation of the reflection lines observed in X-rays of crystals. The key idea of their model was that X-rays would scatter off the individual atoms in

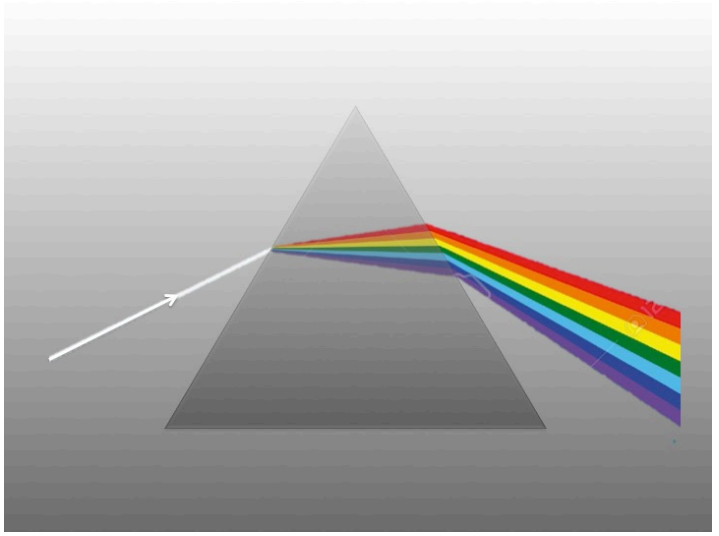


Figure II.3.8: *Color decomposition of white light through a prism.* The refraction of white light by passing through a prism. The propagation speed of light of different colors (frequencies) is different in glass and leads to different amounts of refraction.

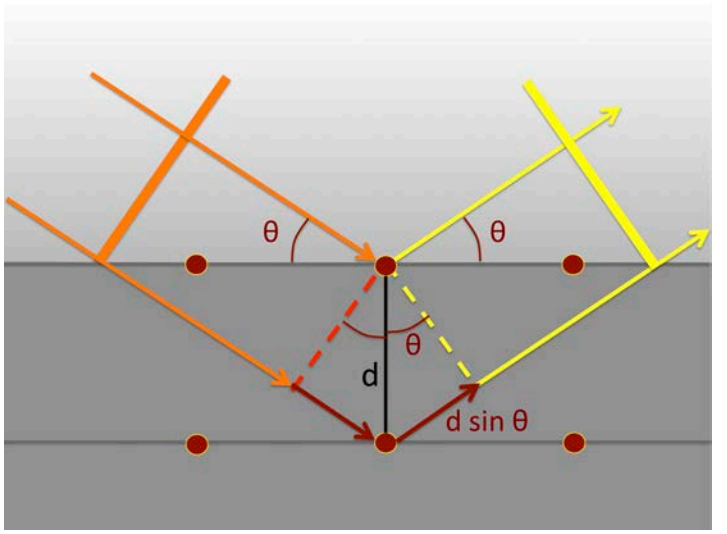


Figure II.3.9: *Bragg reflection.* The crystal consists of equally spaced layers. Two rays from the top two layers are drawn, the path difference between the incoming and outgoing wavefronts of the two paths equals $2\mathrm{d}\sin\theta$, this should equal an integer times the wavelength λ.

subsequent layers of the crystal. The layers in a crystal are equally spaced with a distance d, a distance that is typically about $10^{-10}\,\mathrm{m}$. Requiring radiation with a wavelength comparable to d yields that we need high frequency X-rays indeed. The question then was to derive the condition for constructive interference of incident and reflected waves. Assuming a monochromatic wave incident under an angle θ with the surface of the crystal, the condition follows from Huygens' principle, as is schematically depicted in Figure II.3.9. The path length between the two rays scattering from the two top layers should be proportional to an integer m times the wavelength to obtain constructive interference. The integer m is called the *diffraction order*. This leads to the Bragg formula:

$$2\mathrm{d}\sin\theta = \mathrm{n}\mathrm{m}\lambda\,.$$

This formula is general as long as the particles in the beam are scattered in a spherical fashion from each individual atom in the lattice. In that sense the formula can also be applied to matter waves, in other words, to the scattering of electrons or neutrons from crystal surfaces. By looking at different plane orientations this principle turns into a powerful technique to determine the spatial structure of crystals.

Beamsplitters and polarization

In classical optics it was Newton who in his *Opticks*, published in 1704, introduced the prism to split a beam of light into its different light components (see Figure II.3.8), while Huygens in his monumental *Traité de la lumière*, published in 1690, emphasized the importance of double breaking by 'Icelandic crystal' or calcite, and explained it to a certain extent with his wave theory of light.

These explanations were all based on the idea that different components of 'ordinary' light have different velocities

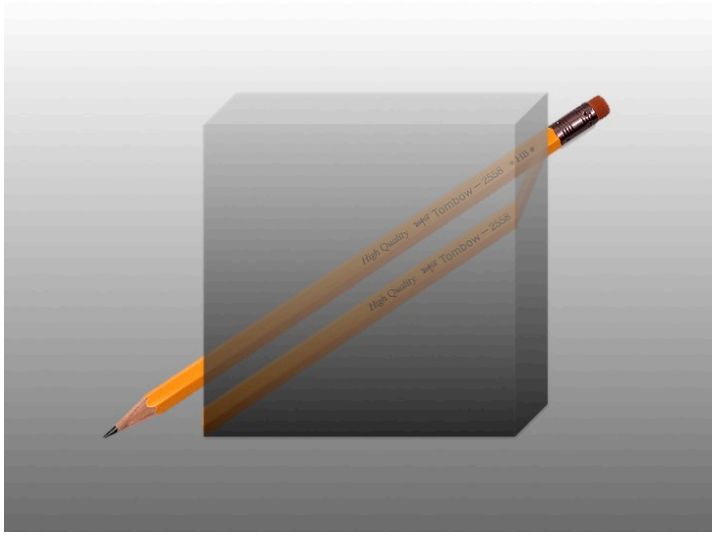


Figure II.3.10: *Icelandic crystal.* Double refraction of light by an Icelandic crystal or calcite.

in various media, and therefore have a different amount of refraction at interfaces between various media. And this is indeed a fundamental ingredient of all beam splitting devices. We should be aware that in the early theories of light that arose in the Enlightenment era through the works of Descartes and later of Huygens and Newton, many properties of light were discovered and these led to the great dispute between the latter two about the particle versus wave-like nature of light. The property of polarization was not really discussed, and understanding the transversal wave nature of light had to wait untill Maxwell identified light as an electromagnetic waves two centuries later.

However it is remarkable to see how tantalizing close Huygens came to discovering the nature of polarization exactly because of his particular emphasis on the phenomenon of bi or double refraction exhibited by light passing through an *Icelandic crystal*, which we have depicted in Figure II.3.10. This phenomenon occurs basically in all transparent anisotropic media. In his treatise he remarks:

> Before finishing the treatise on this Crystal, I will add one more marvelous phenomenon which I discovered after having written all the foregoing. For though I have not been able till now to find its cause, I do not for that reason wish to desist from describing it, in order to give opportunity to others to investigate it. It seems that it will be necessary to make still further suppositions besides those which I have made; but these will not for all that cease to keep their probability after having been confirmed by so many tests.

He then goes on to describe how he studied the properties of light subsequently passing through two crystals and makes the observation that the double refraction does not take place at the second crystal, as is clear from his illustration (see Figure II.3.11). He even goes as far as to observe that the properties of the second refraction depends on the orientation of the crystal. And his humble conclusion reads:

> It seems that one is obliged to conclude that the waves of light, after having passed through the first crystal, acquire a certain form or disposition in virtue of which, when meeting the texture of the second crystal, in certain positions, they can move the two different kinds of matter which serve for the two species of refraction; and when meeting the second crystal in another position are able to move only one of these kinds of matter. But to tell how this occurs, I have hitherto found nothing which satisfies me.

In the following we discuss various cases of how the splitting of a beam, dependent on the polarization state of the particles can be achieved. First we discuss some beam splitters for photons. Next we discuss the case of spin one half particles like electrons, protons and neutrons in a magnetic polarization device like the Stern–Gerlach setup. We also introduce some other devices from which more

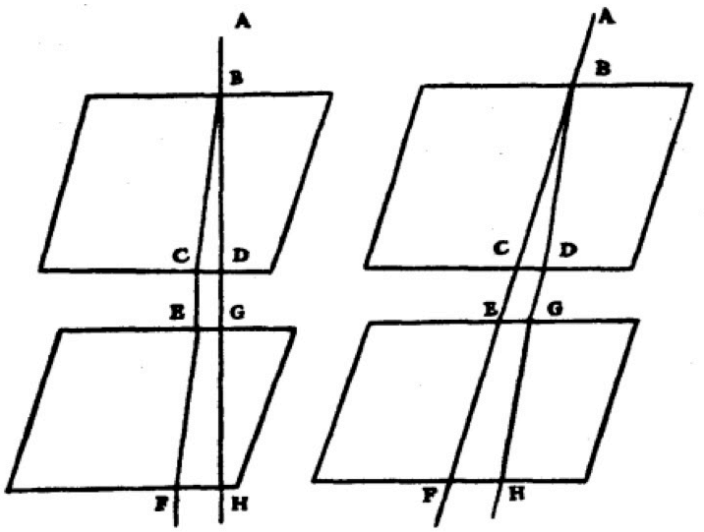


Figure II.3.11: *'A marvelous phenomenon.'* Double refraction of light does not occur in the second crystal. Illustration taken from Huygens' *Treatise on light.*

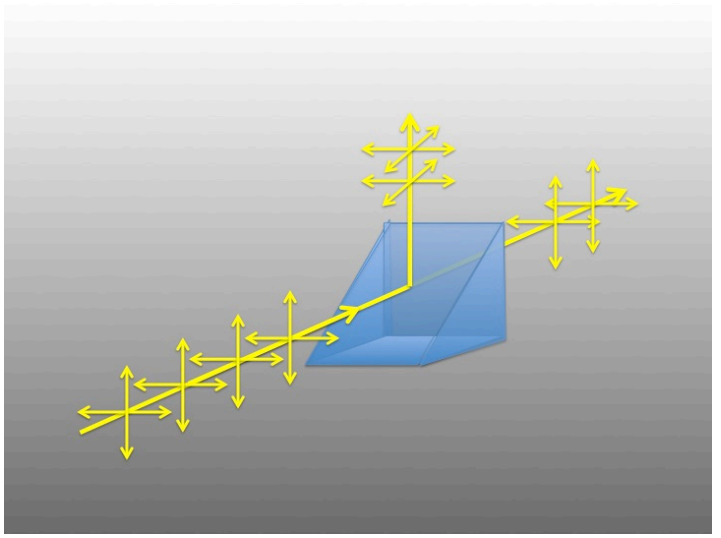


Figure II.3.12: *A half mirror.* A half-silvered mirror reflects half the number of photons in a beam, the other half is transmitted. It is a beam splitter (BS) that is insensitive to the polarization state of the incoming photons.

elaborate interference experiments can be assembled. Together, they form part of the toolkit for many famous experiments that demonstrated how different quantum theory really is, where particles can interfere with themselves, or where certain forms of non-locality (which are strictly forbidden in the classical realm) pertaining to entangled states of particles can be unambiguously demonstrated. This will be our focus in the remainder of this chapter.

Photon polarization: optical beamsplitters

In modern (quantum) optics using monochromatic lasers, many quite stunning experiments have been performed, demonstrating the paradoxical but quantessential features of light and in particular its polarization. In the previous chapter we have already discussed various filters: polarizers on page 293, and wave plates on page 290 through which the polarization states can be selected and/or manipulated. Now we extend the toolset with some beamsplitters much in analogy with the Icelandic crystal. These devices play a crucial role in experiments where properties like particle interference and entanglement can be put to the test.

Clearly, by splitting a beam one obtains two beams which are strictly in phase and therefore offer interesting experimental possibilities.
A first splitting device would be the *half-mirror*, where half the number of photons in the beam gets reflected while the other half gets transmitted. As such this mirror is insensitive to the polarization state of the photons, as we have indicated in Figure II.3.12.
It is also possible to coat the interface with particular chemicals in which case we obtain a *polarizing beam splitter* as depicted in Figure II.3.13; if the incoming beam is unpolarized, the reflected photons are horizontally polarized, while the transmitted ones are vertically polarized. We obviously can rotate the polarizing cube around the incoming

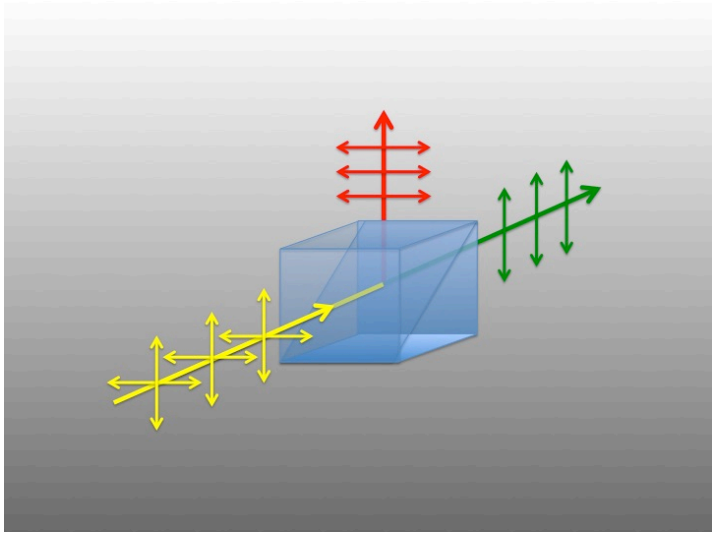


Figure II.3.13: *A polarizing beam splitter (PBS).* This component is sensitive to the polarization state of the photons: the reflected ones are horizontally polarized, the transmitted ones vertically.

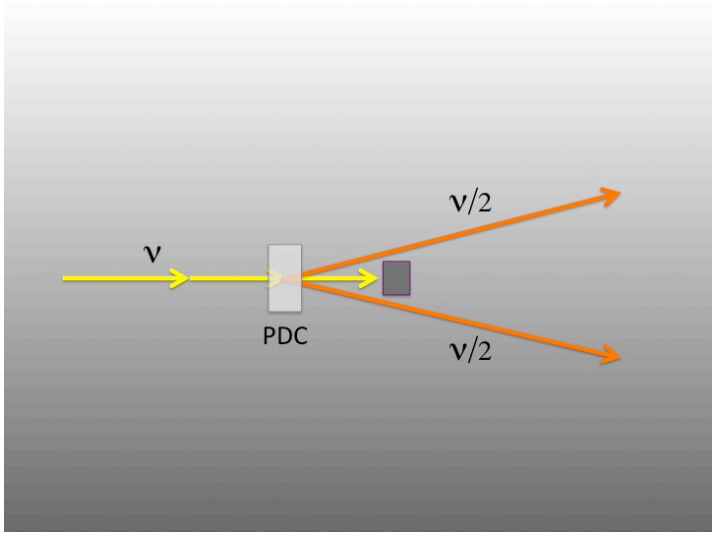


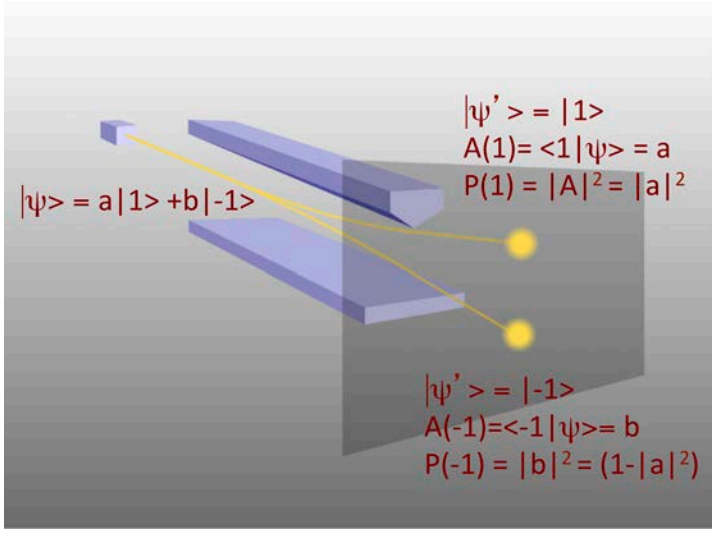
Figure II.3.14: *Two photons out of one.* A parametric down converter (PDC) is a nonlinear crystal where incoming photons may be converted down to two photons with half the energy or frequency. These secondary beams leave the crystal under a small angle with the primary beam. The polarizations of the secondary pair are entangled and can be chosen to be either parallel or orthogonal.

photon momentum vector, generating a split between two other linear polarizations. This device acts much like the anisotropic crystals causing double diffraction like the ones Huygens mentioned. It is also similar to the Stern–Gerlach device to split a beam of spin-$1/2$ particles to which we turn shortly.

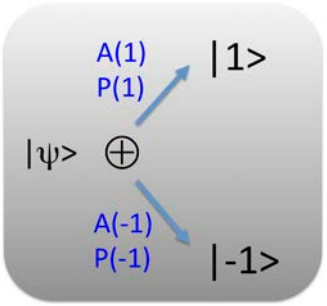
A final device we want to mention is what is called a *parametric down converter*. It is a nonlinear crystal that splits an incoming monochromatic beam of a given frequency f; it splits a fraction of the incoming photons into two photons with half the frequency (or energy). These secondary photons leave the crystal under a small angle with the incoming beam as we have indicated in Figure II.3.14. As we will discuss later, the remarkable property of these secondary pairs is that their polarization states are entangled. Depending on the type of crystal this maybe parallel or orthogonal entanglement, where one speaks of type I or type II down conversion.

Spin polarization: the Stern-Gerlach device

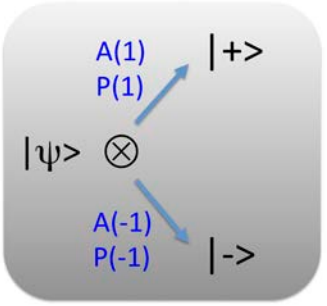
We have illustrated this means of polarizing the spin for various choices of the state $|\psi\rangle$ and observable A being the spin polarization, in the Figures II.2.9. Let us comment on their content. The green circle is the space of normalized quantum states; normally this would be a three-sphere but we have chosen the section where the coefficients α and β are real, so we are left with an ordinary circle in R^2 . We consider two real observables being Z and X and combinations thereof, those have always eigenstates that are lying on the circle. In the diagrams in the figure we see pairs of blue axes. These axes are in the direction of the eigenvectors of A and labeled by their eigenvalue. The blue axes together represent thus the *measurement frame* corresponding to A . Now there are five things to observe: (i) the blue axes have a direction but are not oriented, ex-



(a) Stern–Gerlach experiment: Measurement of spin polarization along z-axis, of a state $|\psi\rangle$. Outcome can only be $+1$ or -1 in units $\hbar/2$ with probabilities depending on the particular state $|\psi\rangle$. The measurement outcome affects the state of the outgoing particle.



(b) Measurement spin polarization along z-axis, of a state $|\psi\rangle$.



(c) Measurement spin polarization along x-axis, of a state $|\psi\rangle$.

Figure II.3.15: *The Stern–Gerlach experiment.* (a): Sending the electron beam through an inhomogeneous magnetic field will split the beam. (b) and (c): Symbolic representation of the Z and X polarizing beam splitters that we will use later on.

pressing the fact that the opposite points $\pm|\psi\rangle$ have the same probabilities. They are indistinguishable by measurement. In other words they only differ by a phase, which in this case a real phase, which can only be -1;
(ii) perhaps it is also surprising that the frames corresponding to Z and X are *not* orthogonal, rather they only make an angle of 45^o , half the expected angle. If we were to turn the polarizer in the $\mathrm{minus}\, z$ direction, thus rotating in *real* space the polarizer in the plane by 180^o , would interchange the eigenvalues and consequently interchange the axes of the measurement frame, which is equivalent to rotating in *state* space by half the angle, in this case 90^o . Saying it yet differently: we have chosen the up and down state vectors of the spin as orthogonal unit vectors. This means that if we rotate the device by φ in ordinary x, y, z-space, then the polarization plane will only rotate by $\varphi/2$ degrees in spinor space, which in the Hilbert space for this system means a rotation by 90^o . That explains why the choice of observable involves fixing two orthogonal axes in state space; it is really a choice of frame rather than selecting a particular direction.
(iii) Once the measurement has been made, one axis of the frame has been singled out, and the wavefunction 'collapses' to a normalized state along that axis. If in the example of Figure II.2.9(d) above, we happen to measure the X eigenvalue $+1$, then the state collapses along the corresponding axis, meaning that we move from the state in Figure II.2.9(d) to the state in Figure II.2.9(c).
This picture indeed allows us to make the projections on the axes which give the probability amplitudes while the measurement outcome labels the axis, and they also tell you what the collapsed state looks like.

Indeed this graphical representation captures some quant- essential features of the measurement process. We will make use of it repeatedly later on.
(iv) The analysis we just presented underscores the subtle meaning of the 'state vector' or wavefunction. Indeed it is important to always keep in mind that it is as much defining a state as it is a probability amplitude, which means a way of encoding probabilities of measurement outcomes of any given observable.
(v) Bearing the previous points in mind there is an additional remark to be made at this point. Did we make a measurement or not necessarily? When we put a screen

Barbie's choice. Let us now rephrase the measurement process in the language of the Barbie on the globe, the representation of spin space we introduced in the figure on page 285 . The geometry is now somewhat different: we first have the spin in a certain state, which means that the Barbie is located at some point on the 2-sphere and pointing her nose in a certain direction of the tangent plane at her location.
Making a measurement amounts to choosing an orientation in the X, Y, Z space, which we can mark as a line through the center and intersecting the unit sphere into two antipodal points on the sphere. The intersection of the positive direction with the sphere corresponds to the positive eigenvalue eigenstate, and the negative intersection corresponds to the negative eigenvalue eigenstate. Indeed the choice of observable determines the eigenstates up to a phase factor. So, staying within the narrative, choosing the orientation of the detector corresponds to installing two inspectors at the corresponding antipodal points on the sphere. These inspectors do not look in any specific direction, they just search around and try to spot the Barbie. Once they have spotted her they both call to her (the sphere is of course transparent – a crystal three-sphere...) and order her to report immediately at their place. Barbie doesn't quite know who of the two to choose, but she makes a choice, it doesn't matter who Barbie chooses as long as the probabilities are in accordance with her little quantum calculation. The inspectors go home and leave her on the spot she happened to choose. That's the state she ends up in, and that was what the measurement was.■

and record the electron hitting the screen we surely have made a measurement of its spin. But you may also imagine an experiment where we do not register (or measure) it explicitly, but think of the experiment as a way to select the initial spin state for some other experiment that makes use of the upper or lower beam. Then it is clear that the Stern Gerlach devise is used as a preparatory device to select an incoming spin state.

And that naturally accommodates the fact that the state alters after a measurement, because the information we gather from the measurement may drastically affect the probabilities. It is not that we as observers play a role, because we may or may not look at the results, it is the interaction that has or has not taken place between the apparatus and the system, which matters.

Interference: double slit experiments

An important property of waves is that if we combine two of them their amplitudes are added together and we get interference: in places where the waves are in phase the combined wave gets a maximal amplitude and where they are out of phase they will compensate resulting in a reduced amplitude.

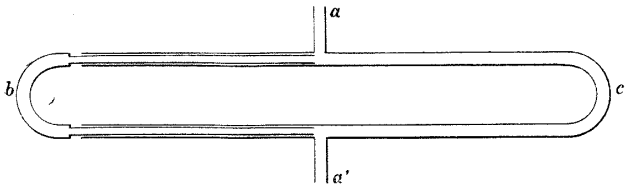


Figure II.3.16: *Interference.* A 'sound' interference experiment, due to Georg Hermann Quincke, which demonstrates the interference of sound waves. Image from a 19th century high school book on physics.

The interference of sound. A simple demonstration of classical interference can be given with the sound experi-

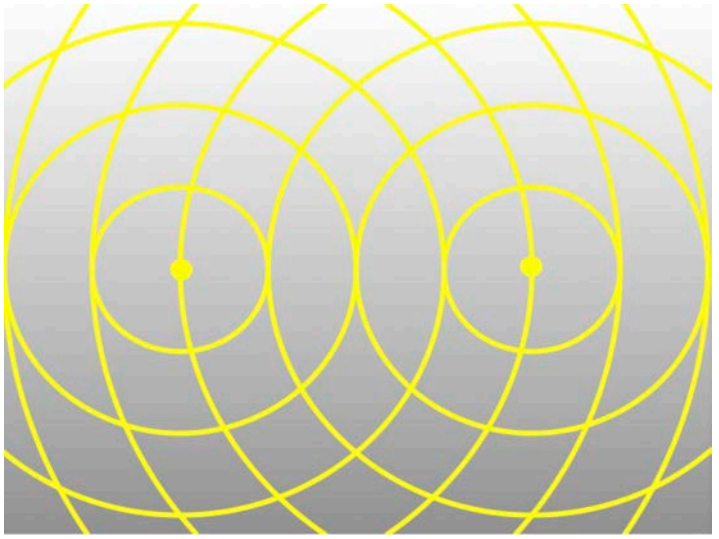


Figure II.3.17: *Two point sources emitting waves.* The two sources are 4 wavelengths apart and are in phase.

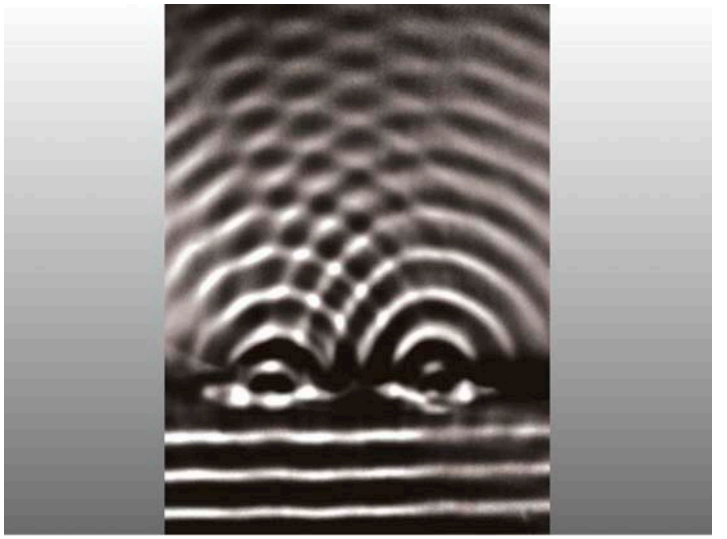


Figure II.3.18: *Water waves.* Two slits act as sources emitting water waves that interfere. This illustrates the geometric constructions displayed in the following figures.

ment devised in the nineteenth century by Georg Hermann Quincke as shown in Figure II.3.16. In the modern guise a tone is generated with small loudspeaker at the point a on top, the sound (air pressure) wave splits and propagates through both the left and right tubes. They come together again at the point a' at the bottom, where the two waves interfere. The difference in length between the left and right paths can be adjusted so as to obtain constructive or destructive interference. In the latter case a microphone positioned at a' would not register any sound. The crucial thing is that the total signal at the microphone is built up from the various amplitudes along the two independent paths and in that sense this is really a kind of double slit experiment.

Wave interference from two point sources. Figure II.3.17 shows two point sources emitting circular waves which are in phase. The two individual wave patterns overlap and will therefore interfere, meaning that at certain points the signals will amplify each other and in other points they will cancel. A new pattern of maxima and minima will develop. In Figure II.3.18 we show the pattern of water waves generated by two point sources that oscillate in phase (almost). The pattern is obtained by literally adding up the amplitudes of the two individual spherical patterns coming from the two slits which act as point sources, incoming are the plain waves from below and this makes that the two sources oscillate in phase. So this is indeed a double slit experiment and we see that the resulting pattern has a number of striking features. We roughly see rays of outward moving waves with indeed an amplitude that varies depending on the angle.

In Figure II.3.19 we give the theoretical reconstruction of the situation combining the two previous figures. In the top half of this figure we could mark the points by the path difference from the two sources (which equals an integer times the wavelength) and then connect the points with equal differences, as we did in Figure II.3.20. What we obtain are the orange colored hyperbolic rays along which the maximal amplitude oscillations propagate upwards. In between we could have drawn the zero amplitude node lines connecting points where the difference is

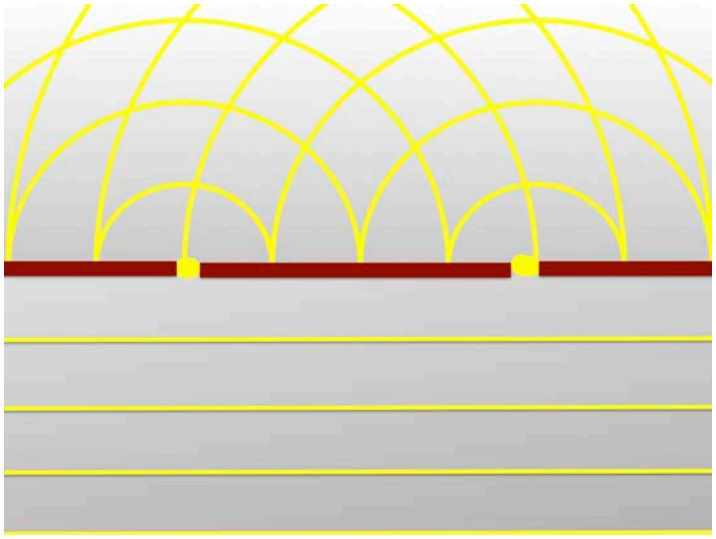


Figure II.3.19: *Double slit interference.* The slits act as sources emitting semicircular waves that will interfere. Compare the pattern with the water wave interference pattern of the previous figure.

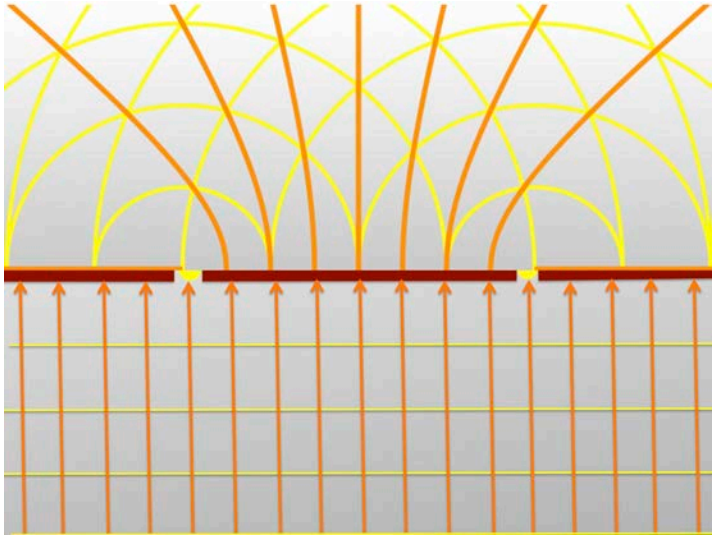


Figure II.3.20: *Rays.* The orange maximal amplitude rays connect points that have distances to the sources which differ by a certain integer times the wavelength.

a half-integral multiple of the wavelength. The pattern of rays that emerges is not entirely obvious, because there is no such thing as 'adding' rays; you add the wave patterns and then construct the resulting ray pattern.

Once we have the pattern of rays we could also draw the new wavefront picture. These correspond to the blue elliptic curves in Figure II.3.21. Note that indeed the rays and wave fronts are orthogonal in any point where they meet. Rays and wavefronts always form what is called two orthogonal families of curves. What you will see is that these wave fronts move outward. So what is the picture along any one of these wave fronts? It crosses a fixed number of maximal amplitude and node rays and these rays stay fixed in time. Therefore we would encounter a one-dimensional *standing wave pattern* along the wave front, and that is what is visible in the water wave picture II.3.18.

The interference of light. In Figure II.3.22 we have depicted the classical experiment of Young in which he showed the interference of the light going through the two slits. It only occurs if both slits are open. If only one slit is open, one gets a single maximum comparable to that of classical particles. The result was fully consistent with Huygens' principle of light propagation following from the wave nature of light. Comparing this experiment with the previous one on sound waves it is clear the sound measurement only corresponds to a single point on the detection screen for the light. Moving the trombone arm on the left of Quincke's device corresponds with moving the light detector up and down the screen, which is necessary to probe the minima and maxima of the interference pattern.

The non-interference of marbles (classical particles). In Figure II.3.23 we see a source shooting particles (say, marbles) in all forward directions. Most get absorbed by the screen but if they are directed to one of the two slits, the particles can get through. If we count the number of particles hitting the detector screen, we typically get a distribution with two single maxima as indicated in Figure II.3.23. This is exactly what one would expect: there is no inter-

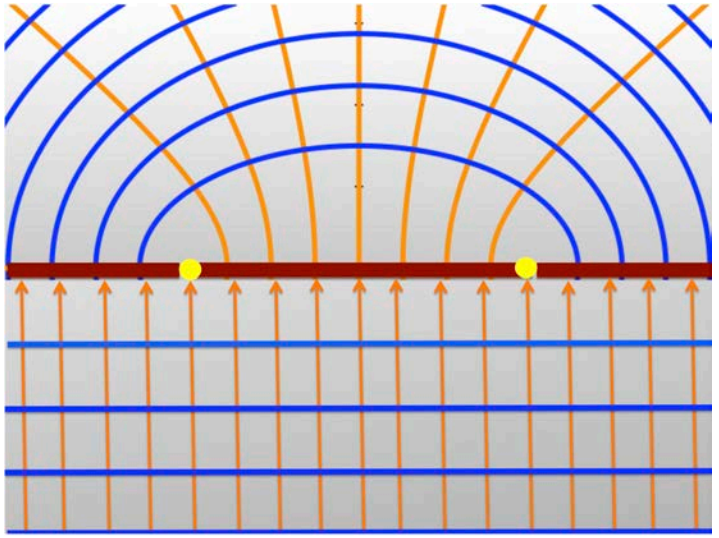


Figure II.3.21: *Wave pattern.* If we draw the elliptical curves orthogonal to the rays, we do not get the familiar wave fronts where the phase difference is constant. Along the curves one obtains a standing wave with varying amplitude and wavelength.

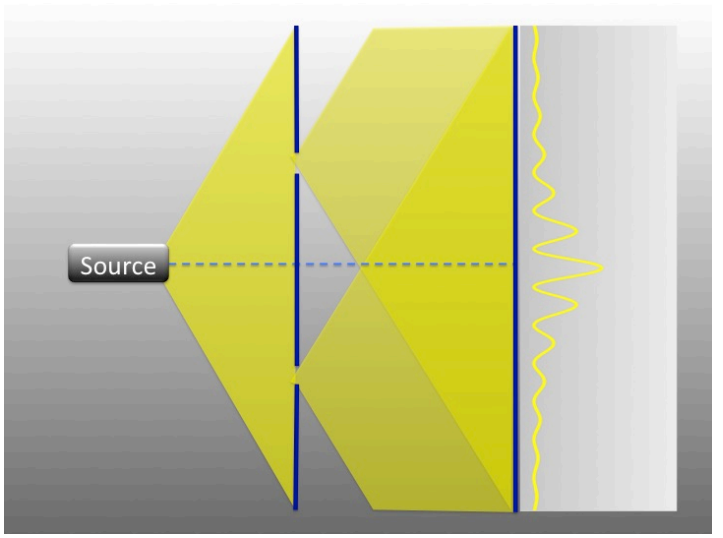


Figure II.3.22: *Young's experiment.* The double slit experiment for light as performed by Young to demonstrate the wave character of light, thereby confirming Huygens' theory of light. On the right the varying intensity of the light on the screen due to the interference.

ference of marbles, let alone that a marble would interfere with itself.

The self-interference of a quantum particle. In Figure II.3.24 we have sketched what happens with a beam of quantum particles such as electrons or protons or neutrons when they hit a screen with two narrow slits. The quantessence is that it does not repeat the pattern of the classical particles of Figure II.3.23 but rather that of light depicted in Figure II.3.22. This fundamental experiment demonstrates the wavelike nature of particles in the quantum domain. The most remarkable, really quantessential aspect of this behavior is that the phenomenon is *not* a consequence of different particles in the beam interfering with each other. This would make it a collective phenomenon, but no, the truly remarkable fact is that if you shoot the electrons one by one, then the interference pattern would slowly build up as is shown in Figure II.3.25. This implies that each electron somehow interferes with itself, and one has to conclude that each electron has 'knowledge' of the probability distribution as a whole.

This is indeed the case in quantum physics, as the wavefunction of the particle is exactly the probability amplitude for finding it in any place at any time. Alternatively you may say that in quantum theory you could calculate the probability for distinct paths from the beginning to any endpoint on the screen separately, then the total amplitude from the beginning to that given endpoint is the sum of those amplitudes. It is the linear superposition principle in a different guise. Let us go one step further and assume that the state $\psi_1(x)$ describes the wavefunction for the configuration with only the left slit open, and $\psi_2(x)$ with only the right slit. The (normalized) wavefunction for the experiment with both slits open would then correspond to $\psi(x) = (\psi_1(x) + \psi_2(x))/\sqrt{2}$, as we just have to add the amplitudes. The probability of finding the particle on a screen behind the slits is then not the same as the sum of the probabilities of the individual left and right slit experi-

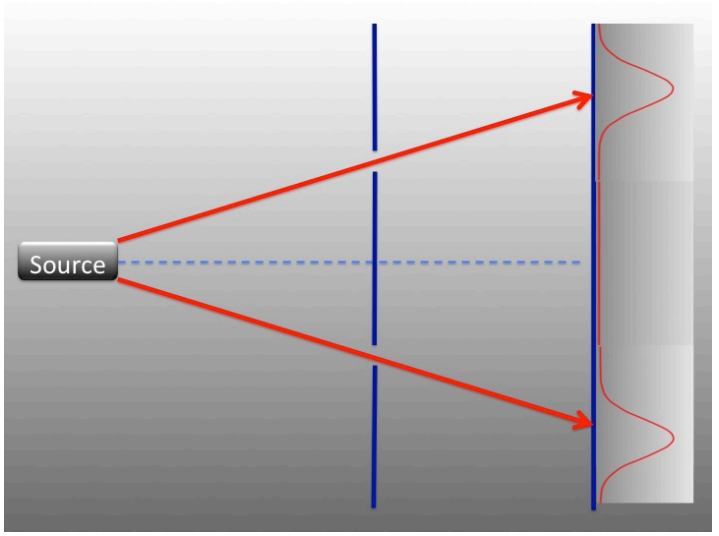


Figure II.3.23: *Marbles don't interfere.* In the double slit experiment with classical particles, the number density of particles hitting the detector screen has two separate maxima and there is no interference.

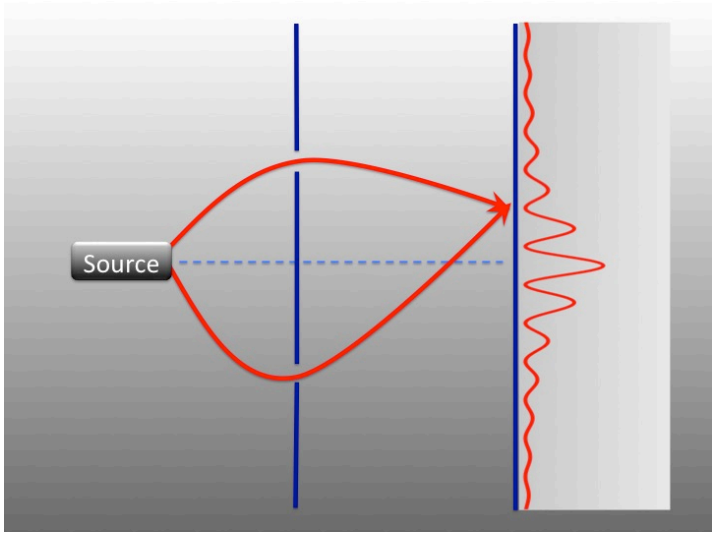


Figure II.3.24: *Electrons are not like marbles.* The double slit experiment showing two conceivable paths that a quantum particle like the electron may have taken. The variation in the intensity pattern on the screen demonstrates the wave nature of quantum particles.

ments, because squaring the total amplitude, yields

$$p(x) = \frac{1}{2}(p_1(x) + p_2(x)) + I(x) \,,$$

where the interference term $I(x)$ is defined as

$$I(x) = \frac{1}{2}\left(\psi_1^*(x)\psi_2(x) + \psi_1(x)\psi_2^*(x)\right) . \qquad \text{(II.3.1)}$$

This is basically the one-particle *quantum interference* effect, a direct consequence of the particle-wave duality in quantum physics.

In talking about quantum interference we should appreciate that a single particle is described by a wave pattern that may or may not be considered to be composed of different components, and therefore a particle can 'interfere with itself' because of the linear superposition principle. And that is what makes quantum interference a truly quantessential phenomenon.

At this point there is an additional remark I would like to make. The question whether or not an interference pattern for the quantum particle will appear depends in a subtle way on what the experimental setup is. For example, look at the experiment of Figure II.3.26, where we have introduced a source which emits pairs of entangled particles; and particle 2 goes to the left and may or may not be detected, while particle 1 goes to the right in the direction of the double slit. The question is whether or not we will see an interference pattern as in Figure II.3.24. The answer is, that whether we will or will not see interference depends on the state of particle 2, *irrespective* of whether we actually measure particle 2! It is the mere *possibility* of identifying the path that particle 1 has taken that destroys the interference pattern. The state of the entangled particles is basically,

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|\mathrm{red}_1\rangle|\mathrm{red}_2\rangle + |\mathrm{green}_1\rangle|\mathrm{green}_2\rangle \,. \qquad \text{(II.3.2)}$$

The interference term for particle 1 would come from the red-green cross term appearing in $\langle\psi|\psi\rangle$ evaluated along

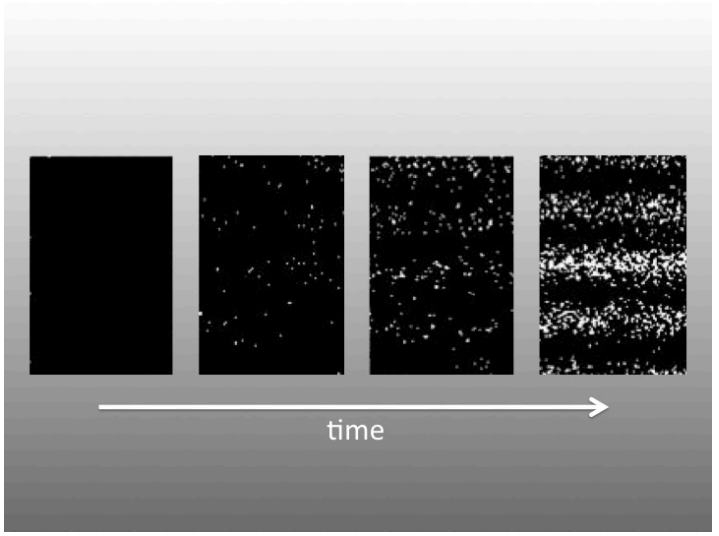


Figure II.3.25: *How particles make a wave pattern.* Buildup of the interference pattern of Figure II.3.24, from the successive hits of single particles (like electrons) on the screen.

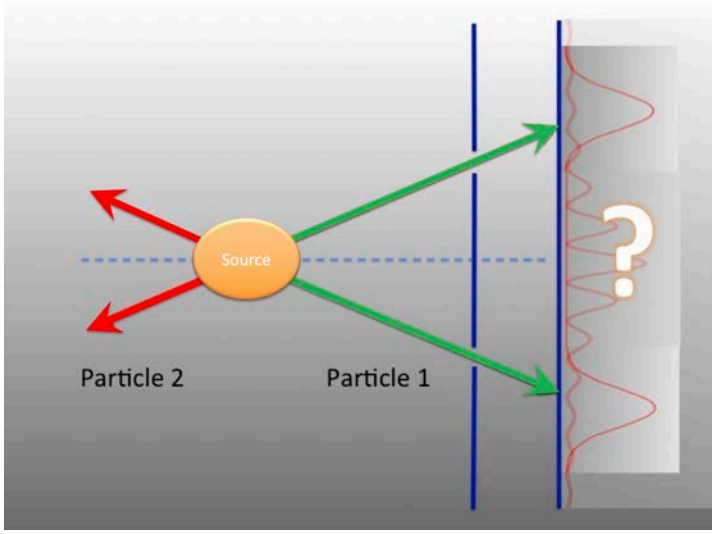


Figure II.3.26: *'Which path' information.* No interference of particle 1 (moving to the right) if it is entangled with particle 2 and thus a path identification would be possible *in principle* by measurement of particle 2.

the screen:

$$\mathrm{I_{rg}} = |\langle \mathrm{red_1}|\mathrm{green_1}\rangle|\,|\langle \mathrm{red_2}|\mathrm{green_2}\rangle|\,,$$

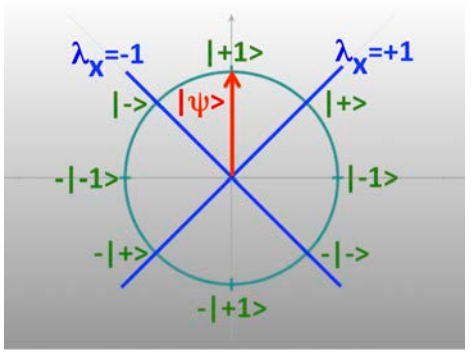
and this term containing the self interference of particle 1 in the first factor will vanish if the second factor for particle 2 vanishes because the $|\mathrm{red_2}\rangle$ and $|\mathrm{green_2}\rangle$ states are orthogonal. Orthogonality here means that they have no overlap: $\langle \mathrm{red_2}|\mathrm{x}\rangle\langle \mathrm{x}|\mathrm{green_2}\rangle = 0$ for all x. If they are not, (some) interference will result, but as you see this really depends on the actual setup of the experiment. As entanglement with the environment can easily take place, sufficient care has to be taken if one wants to demonstrate quantum interference effects. Physicists have gone one step further by investigating the effect of *erasing* the tracking information of particle 2, and they have shown that if you succeed in constructing a quantum eraser in your setup, the interference pattern will emerge. These in-between cases have been investigated in many different types of experiments. We will discuss one such experiment for photons shortly.

It is the non-commutativity of observables that gives rise to the intricacies in the quantum theory of measurement. The predictions of quantum mechanics are intrinsically probabilistic yet the theory is essentially different from classical probability theory. On the one hand it is clear that a given operator defines a probability measure on Hilbert space; however as the operators are non-commuting (like matrices) one is dealing with a non-commutative probability theory, and complementary measures.

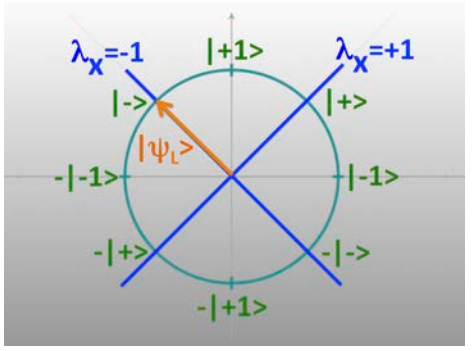
A basic interference experiment

We have illustrated the schematic of a typical quantum interference experiment in Figure II.3.27 which compares two different states and their superposition in the familiar spin or qubit system.

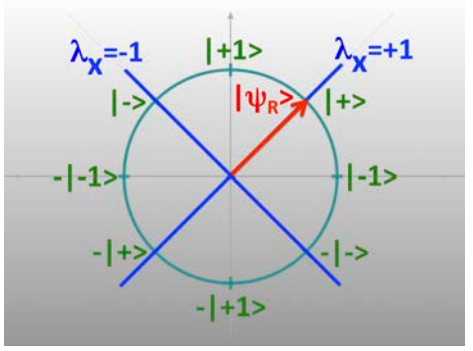
In the top figure (a) we have a beam incoming identically prepared spins that goes through a polarizer in the x di-



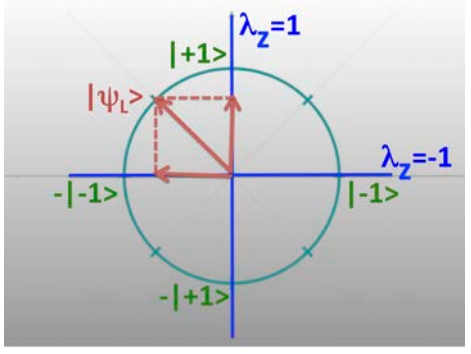
(a) X polarizer and beamsplitter.



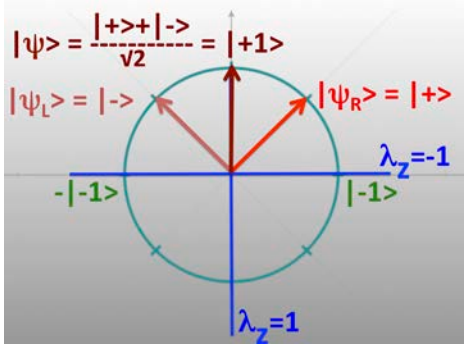
(b) $\mathrm{X} = -1$ polarized in left channel.



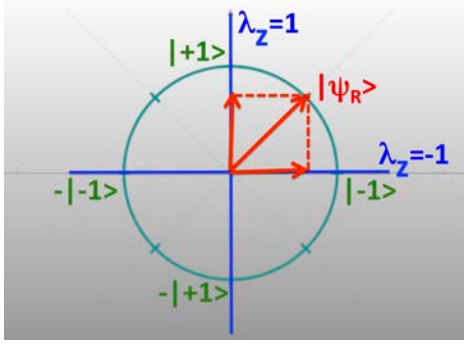
(c) $\mathrm{X} = +1$ polarized in right channel.



(d) *Experiment 1:* Measurement Z after blocking right channel: $\mathrm{p}(+1) = \mathrm{p}(-1) = \frac{1}{2}$.



(e) *Experiment 3:* Measurement Z of left-right interference: $\mathrm{p}(+1) = 1$ and $\mathrm{p}(-1) = 0$.



(f) *Experiment 2:* Measurement Z after blocking left channel: $\mathrm{p}(+1) = \mathrm{p}(-1) = \frac{1}{2}$.

Figure II.3.27: *Three experiments.* Schematic of a typical quantum interference experiment. Adding the red amplitudes of left (d) and right (f) gives the purple amplitude of (e). The corresponding probabilities *do not* add.

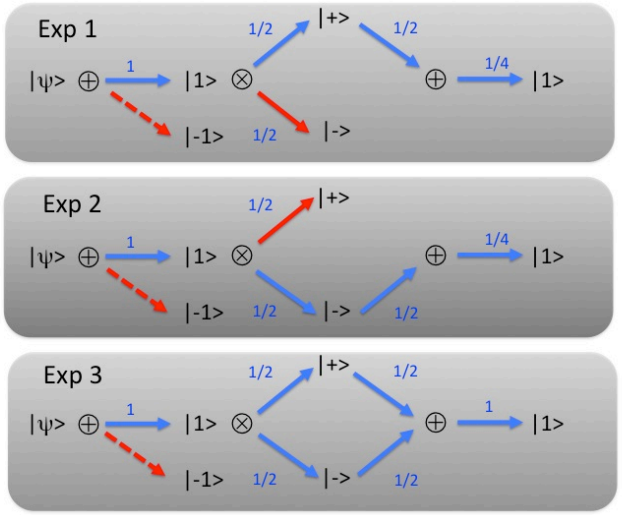


Figure II.3.28: *Adding probabilities.* A different schematic view of the three interference experiments of Figure II.3.27 using the symbolic notation of Figure II.3.15. Adding the final probabilities of the first and second experiment does not give the probability of the third experiment.

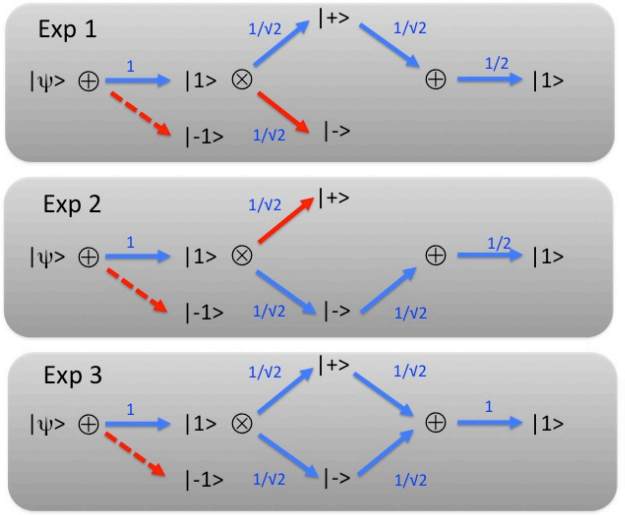


Figure II.3.29: *Adding probability amplitudes.* The same schematic view of the three interference experiments as the previous Figure II.3.28. In this figure we give the probability amplitudes and now one finds that adding the total amplitudes of the first and second experiment *does* give the amplitude of the third experiment.

rection and is split into a left and right channel with opposite polarizations as shown in the two middle row figures (b) and (c). It is important to bear in mind that what we say next applies to each particle individually. The beam is just there to allow us to do a series of repeated measurements in each setup. In the bottom row we have depicted the probabilities for three distinct experimental setups, all measuring the Z polarization indicated by the blue frame. In figure (d) we give the situation if the right channel were blocked where we have $|\psi\rangle = |\psi_L\rangle = |-\rangle$ corresponding to the purple state vector, yielding equal probabilities to measure plus or minus one: $p_L(+1) = p_L(-1) = \frac{1}{2}$. Similarly in Figure (f) on the right we have blocked the left channel, giving $|\psi\rangle = |\psi_R\rangle = |+\rangle$, corresponding to the red state vector in the figure, and we obtain once more $p_R(+1) = p_R(-1) = \frac{1}{2}$. Finally in the middle experiment of figure (e) we have both channels interfere. Adding the probability amplitudes in red of (d) and (f) yields the amplitudes in purple of (e). Now we have to consider the (normalized) superposition $|\psi\rangle = \frac{1}{\sqrt{2}}(|\psi_L\rangle + |\psi_R\rangle) = |1\rangle$ corresponding to the purple arrow in figure (e), so that the probability distribution becomes $p(+1) = 1$ and $p(-1) = 0$. This is notably different from the sum of the probabilities of cases (d) and (f) which would give $\tilde{p}(\pm 1) \equiv \frac{1}{2}(p_L(\pm 1) + p_R(\pm 1))$ yielding once more $\tilde{p}(+1) = \tilde{p}(-1) = \frac{1}{2}$. The differences between $\tilde{p}(\pm 1)$ and $p(\pm 1)$ are indeed due to the interference terms $I(+1) = +\frac{1}{2}$ and $I(-1) = -\frac{1}{2}$.

In Figures II.3.28 and II.3.29 we present an alternative visualization of the same three experiments using the symbols $\oplus$ and $\otimes$ introduced in Figure II.3.15 for the polarizer settings. The left three panels give the probabilities and one sees that they don't add up, while in the right three panels we give the amplitudes and one sees that they do add up. Confirming our expectations for the interference of the spin polarizations.

A delayed choice experiment

A modern and clean quantum incarnation of the canonical double slit experiment is the interference experiment using a so-called Mach–Zender interferometer. In such a device the self-interference of quantum particles/waves, and in particular photons, can be beautifully demonstrated. This setup is also called a *'delayed choice experiment'* after a *gedanken* proposal of John Archibald Wheeler, or a *'which-way experiment'*. The delayed choice refers to the fact that the decision which experiment one is going to do is taken *after* the incoming particles have gone through the first polarizer thereby having chosen one of the two paths or both. In this clever setup the device randomly chooses between:
(i) a 'which way' experiment where one identifies the path which the particle has chosen and thus no interference will take place, or
(ii) a mode where the information on 'which way' is erased and one expects interference.

In Figure II.3.30 I have sketched the schematic of such an experiment[1] by the French group of Alain Aspect, who has pioneered this type of experiments. It consists of two components, first an input part on the left where the polarizations get split. Next the photon travels over a considerable distance of maybe 50 meters (but recently distance of kilometers have been achieved). Finally the photon enters the output part (on the right) where one measures whether the photon has interfered with itself or not. The two components are space-like separated,[2] so that there can be no causal relation between the decision taken in the output part and preparation of the photon in the input part.

Single photons enter the interferometer on the left where they go through a polarizing beamsplitter. The horizontally polarized component goes up and the vertically polarized goes straight through. Reflection by the mirrors does not change the polarization. Then the signal travels some distance. The $\lambda/2$ plate with its axis under 45^o, flips the horizontal and vertical polarizations. This is necessary to allow for the beams to be joined by the second beamsplitter. They traverse the reversed path, so in fact the second splitter acts like a 'joiner'. By tilting the 'joiner' one can also introduce a phase difference φ between the vertical and horizontal component, where the vertical amplitude becomes $e^{i\varphi/2}$ and the horizontal $e^{-i\varphi/2}$. The further encounter depends on the random number generator (RNG) which decides on whether or not to effectively insert another $\lambda/2$ wave plate.

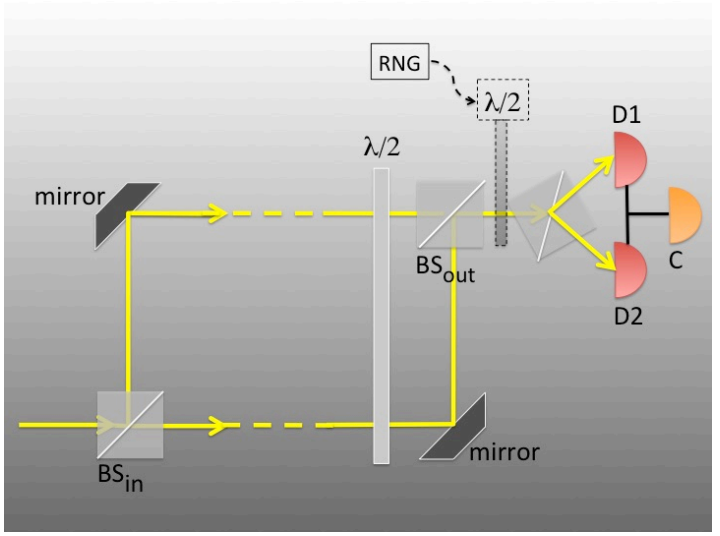


Figure II.3.30: *Delayed choice.* A Mach–Zender quantum interference device, involving two polarizing beamsplitters of the type shown in Figure II.3.13, which demonstrates the quantum interference of photons.

Let us first assume the plate is *not* inserted, then the photon reaches another beam splitting prism that sends the horizontal polarization up to detector D_1 and the vertical polarization down to detector D_2. Furthermore, there is a device that determines whether the detectors 1 and 2 fire

[1] V.Jacques et al., Science, Vol 315 (2007).

[2] Space-like separated means that the output component is outside the future and past light cones of the input component.

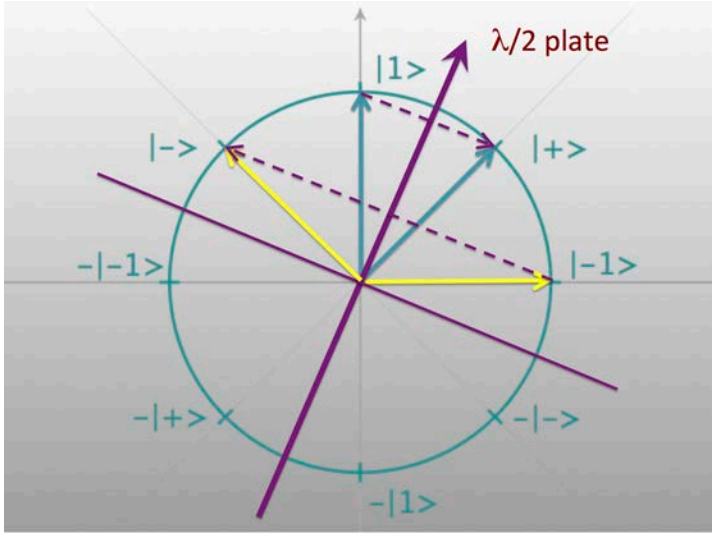


Figure II.3.31: *The* $\lambda/2$ *wave plate.* Effect of the $\lambda/2$ wave with its principle axis under 22.5° with the vertical line. The component orthogonal to the principal axis changes sign (phase= -1). The result is that $|v\rangle = |1\rangle \rightarrow |+\rangle$ and $|h\rangle = |-1\rangle \rightarrow |-\rangle$.

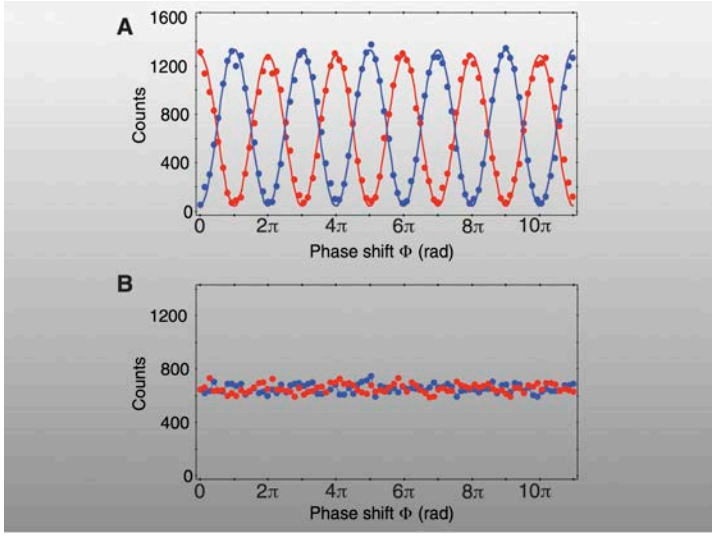


Figure II.3.32: *Single photon interference.* The counts in the detectors D_2 (red) and D_1 (blue), with and without interference. The top graph gives the count with the wave plate in the output channel and the bottom graph without. Taken from V. Jacques et al., Science, (2007).

simultaneously. So the beauty of the setup is of course that all the counts in the detectors are recorded as well as the random time series for the presence of the second plate, and then *a posteriori* one calculates what has happened. Clearly in this mode, the polarization of the photon entering the prism carries the information about which path the photon has taken. The D_1 detects only the photons that came along the lower path, and D_2 detects only the photons that took the upper path. And indeed no interference is observed as is clear from the lower graph in Figure II.3.32. The amplitudes do not add up, and the probabilities are 1/2 and independent of the phase φ . The punchline here is that the whole setup in this mode just 'measures' which path the photon has taken. And knowing that path the photon is just a particle and no interference is to be expected.

In the other mode of the interferometer, an additional $\lambda/2$ wave plate with its axis under an angle of 22.5° is inserted. This has the effect that the polarizations are flipped as Figure II.3.31 illustrates, so that Figure II.3.31 so that $|1\rangle \rightarrow |+\rangle$ and $|-1\rangle \rightarrow |-\rangle$. The important thing is that when the photon enters the final prism the components of different paths will mix again, the amplitudes will add and interference will occur. This is of course assuming the photon took both paths, which is what quantum theory predicts.

So the vertical and horizontal amplitudes become:

$$\alpha_v = \frac{1}{2}(e^{i\varphi/2} + e^{-i\varphi/2}) = \cos\frac{\varphi}{2}$$
$$\alpha_h = \frac{1}{2}(e^{i\varphi/2} - e^{-i\varphi/2}) = i\sin\frac{\varphi}{2}\,.$$

Thus, the probability for counts in D_2 becomes $\cos^2(\varphi/2) = \frac{1}{2}(1+\cos\varphi)$ and that for counts in D_1 equals $\sin^2(\varphi/2) = \frac{1}{2}(1-\cos\varphi)$. And this prediction is beautifully confirmed by the data plotted in the top graph of Figure II.3.32. A single photon interferes with itself, something more quant-essential is hard to imagine.

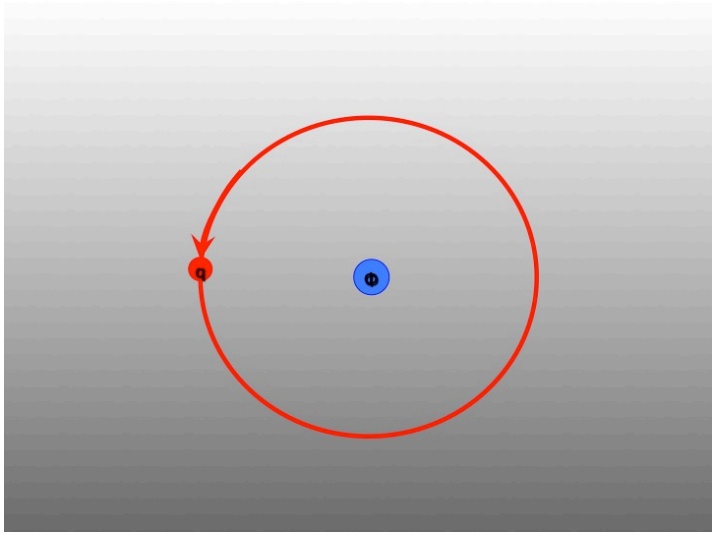


Figure II.3.33: *The Aharonov–Bohm phase factor.* If a charge q encircles a magnetic flux Φ, the quantum state of the particle will acquire a phase factor $W = \exp iq\Phi/\hbar c$.

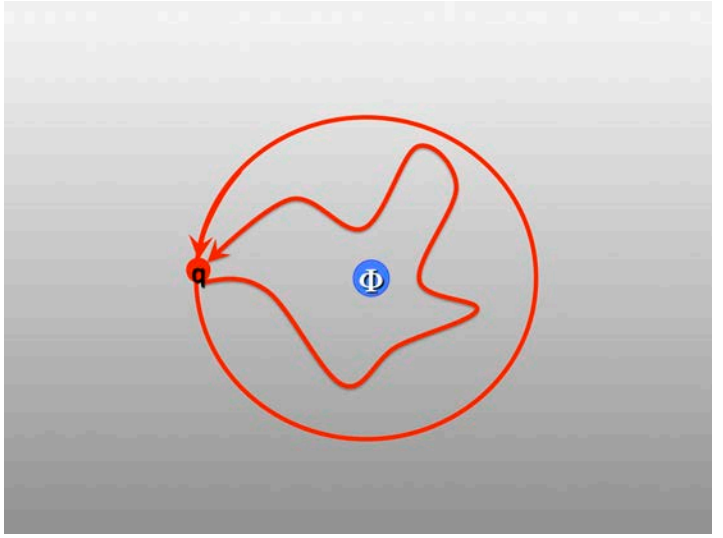


Figure II.3.34: *Path-independence.* The phase factor does not change under deformations of the path, as long as the region in between the paths is free of magnetic fields.

The Aharonov-Bohm phase.

Relative phase factors are all-important in quantum theory and lead to quantessential observable phenomena.

One important example that comes back in many guises is called the *Aharonov–Bohm phase-factor*. The corresponding effect is caused by inserting magnetic flux filament in the one electron double-slit interference experiment. The extra phase that results is due to the line integral of the gauge potential $\mathbf{A}$ along a closed loop, which we introduced already in the section on classical electrodynamics of Volume I in equation (I.1.52) and Figure I.1.27.

Let us recall that if we are in a medium where there is some electromagnetic potential and I have a charge q which I move along a path γ from $\mathbf{x}_0$ to $\mathbf{x}_1$, then the state vector or the wavefunction for that matter will be transformed by a phase factor:

$$\psi(\mathbf{x}_1) = W(\gamma;\mathbf{x}_1,\mathbf{x}_0)\,\psi(\mathbf{x}_0)\,, \tag{II.3.3a}$$

$$W(\gamma;\mathbf{x}_1,\mathbf{x}_0) = \exp\left(i\frac{q}{\hbar c}\int_{\mathbf{x}_0}^{\mathbf{x}_1}\mathbf{A}\cdot\mathbf{dl}\right). \tag{II.3.3b}$$

Here in the integral you take at every point along the path the component of the vector potential directed along the path. The outcome will in general depend on which path you choose. This phase factor is an interesting object, and we should pause for a moment to understand it better.

Firstly note that it is what we call 'non-local,' and under a gauge transformation $U(\mathbf{x})$ it transforms like

$$W(\gamma;\mathbf{x}_1,\mathbf{x}_0) \to U(\mathbf{x}_1)W(\mathbf{x}_1,\mathbf{x}_0)U^\dagger(\mathbf{x}_0).$$

If we close the loop, then the phase-factor becomes gauge invariant, because we get $U^\dagger(\mathbf{x}_0)U(\mathbf{x}_0) = U^{-1}U = 1$, the transformations act at the same point and therefore cancel out.

What does this non-local gauge invariant quantity mean? To understand that we go back to classical electrodynamics, and you have the simple property called Stokes' law, which tells us that if you calculate the line integral of A around a closed loop γ, then you get the magnetic flux through (any) two-dimensional surface bounded by the loop. So this means that the loop operator W_γ 'measures' the magnetic flux:

$$W_\gamma(q,\Phi) = e^{iq\Phi/\hbar c}\,,$$

which is indeed a gauge invariant quantity, as it should be. Let us now go to a two-dimensional situation to simplify the picture, and imagine we have a well-defined narrow magnetic flux tube piercing through the surface as in Figure II.3.33. If we adiabatically move a charge around the flux Φ the state will change according to,

$$|q,\Phi\rangle \to W_\gamma(q,\Phi)|q,\Phi\rangle\,.$$

In Figure II.3.34 we show the effect of deforming the contour or loop doesn't affect the outcome as long as we do not cross magnetic flux lines. In field free regions you can deform the loop arbitrarily. Also, if you first go one way around the flux and you subsequently go back around some other loop encircling the flux in the opposite way, the net effect will be zero.

The beauty of this story is that one can directly measure this gauge invariant phase factor W in a one particle quantum interference experiment. It is called the Aharonov–Bohm effect, after the two theorists who proposed it in 1959 with reference to earlier work by Ehrenberg and Siday.[3] The setup of the experiment is given in Figure II.3.35. The gauge and path independent extra phase factor W_γ appears as a relative phase factor between the ψ_1 and ψ_2 factors in the interference term defined in equation (II.3.1), causing the observed shift of the interference pattern shown in the figure.

[3] And maybe this credential ambiguity explains why there was no Nobel prize awarded for this fundamental effect.

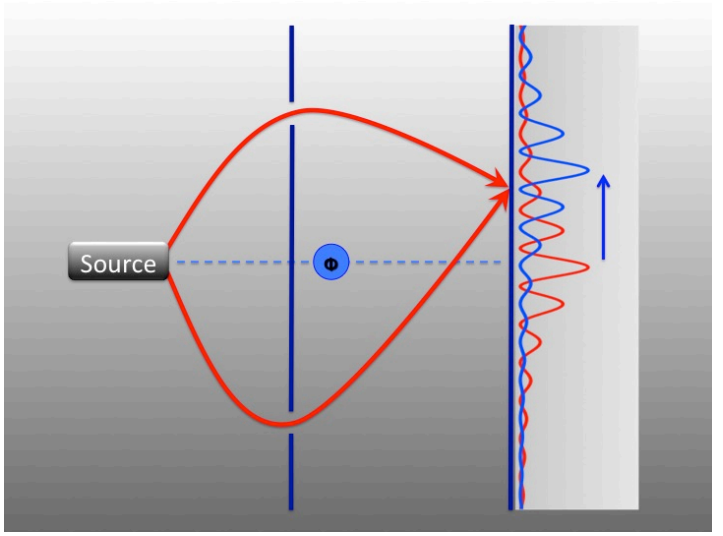


Figure II.3.35: *Path-independence.* The presence of a magnetic flux filament between the slits causes an extra phase difference between the two paths. This leads to a shift of the interference pattern from 'red' to 'blue' as indicated.

Phase shift due to magnetic flux. 

Let us find out how this happens. We start with the free particle Hamiltonian and then include the coupling to the electromagnetic field through the vector potential $\mathbf{A}$, as we did in equation (I.1.44). This amounts to replacing the derivatives $\boldsymbol{\nabla}$ by covariant ones $\mathbf{D} \equiv \boldsymbol{\nabla} + iq\mathbf{A}/\hbar c$:

$$H = -\frac{\hbar^2}{2m}\boldsymbol{\nabla}^2 \quad \to H = -\frac{\hbar^2}{2m}\mathbf{D}^2\,.$$

Suppose we have solved the problem with $\mathbf{A} = 0$ corresponding to $\psi_1(\mathbf{x})$ and $\psi_2(\mathbf{x})$. We want to find out what changes if we take $\mathbf{A} \neq 0$. Consider the covariant derivative working on any function, then we have the following equality:

$$\mathbf{D}\psi^A(\mathbf{x}) = \boldsymbol{\nabla}\Big(\exp\big(i\frac{q}{\hbar c}\int_{x_0}^{x_1}\mathbf{A}d\mathbf{x}\big)\,\psi^A(\mathbf{x})\Big)\,,$$

The only way the coupling to the $\mathbf{A}$ field manifests itself is through the phase factor W. In other words the solutions are linked as follows:

$$\psi_i^A(\mathbf{x}) = W^*(\gamma_i;\mathbf{x},\mathbf{x}_0)\psi_i(\mathbf{x}) \quad ; \quad i = 1,2\,.$$

The phase factor looks awkward in that there at once appears a point $\mathbf{x}_0$ and the line integral along a path γ_i from $\mathbf{x}_0$ to $\mathbf{x}_1$. But the identity holds for *any* choice of $\mathbf{x}_0$ and may depend on γ, as will become clear.

Now return to the interference term $I(\mathbf{x})$ defined by equation (II.3.1). One chooses for $\mathbf{x}_0$ the position of the source, and for ψ_1^A the path γ_1 has to be chosen to pass through the first slit and for ψ_2^A a path γ_2 through the second slit. Then the first term of $I(\mathbf{x})$ involves the product:

$$\begin{aligned}&\psi_1^{A*}(\mathbf{x})\psi_2^A(\mathbf{x})\\&= W(\gamma_1,\mathbf{x},\mathbf{x}_0)W^*(\gamma_2;\mathbf{x},\mathbf{x}_0)e^{i(\beta_2(\mathbf{x})-\beta_1(\mathbf{x}))}\,|\psi_1(\mathbf{x})||\psi_2(\mathbf{x})|\,,\end{aligned}$$

where the the $\beta_i(\mathbf{x})$ are the phases of $A = 0$ solution. Note that $W^*(\gamma;\mathbf{x},\mathbf{x}_0) = W(\gamma;\mathbf{x}_0,\mathbf{x})$, in other words the conjugation reverses the path, but then the product of the two W factors yields a closed path through both slits encircling the magnetic flux giving the overall phase factor $W_\gamma(q,\Phi) = e^{iq\Phi/\hbar c}$. Putting it all together we obtain:

$$I(\mathbf{x}) = \cos\big(\frac{q\Phi}{\hbar c} + \beta(\mathbf{x})\big)\,|\psi_1(\mathbf{x})||\psi_2(\mathbf{x})|\,,$$

with $\beta(\mathbf{x}) \equiv \beta_2(\mathbf{x}) - \beta_1(\mathbf{x})$.

What this calculation shows is that the position dependent phase $\beta(\mathbf{x})$ corresponding to the $\mathbf{A} = 0$ gets shifted by an amount proportional to the flux-charge product. This shift is constant; it does not depend on where you are, which means that the interference pattern generated by $\beta(\mathbf{x})$ gets shifted as a whole, as we have indicated in Figure II.3.35. We will return to these Aharonov–Bohm phases on page 416 of Chapter II.5, where we talk about exotic particle spin and statistics properties in two dimensions. ■

Why is this an important effect? This experiment shows a really interesting aspect of electrodynamics. The electrons in this experiment are shielded from the flux. They only travel through regions of space where both the electric and the magnetic fields $\mathbf{E}$ and $\mathbf{B}$ are strictly zero. The vector potential $\mathbf{A}$ is non-zero but it is a gauge dependent field and therefore not a local physical observable like the other fields. In fact locally it is a gauge transform of the vacuum, in other words locally the gauge potential can always be gauged to zero! And yet, there is an observable effect! The clue is that there is this subtle *nonlocal gauge invariant observable* which involves only the vector potential, namely the loop integral, its value if non-zero cannot be gauged away. This means that if you would like to transform the gauge field to zero everywhere that transformation would *not* be single valued and therefore not be a proper gauge transformation. It is this gauge invariant observable that is measured in this quantessential experiment.

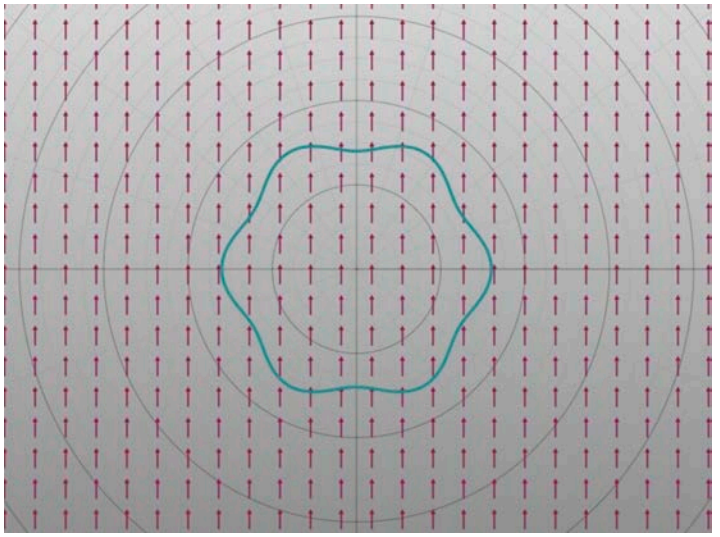


Figure II.3.36: *Super phase (A).* Phase of the superconducting condensate. This is the ground state with the trivial constant phase equal one. This configuration has winding number $n = 0$.

Flux quantization in a superconductor. Let me point out another crucial 'application' of this argument in the context of superconductors, in particular type II superconductors. The defining property of superconductors is that their resistance is zero. If you were to move a piece of superconducting material in a magnetic field, super currents would

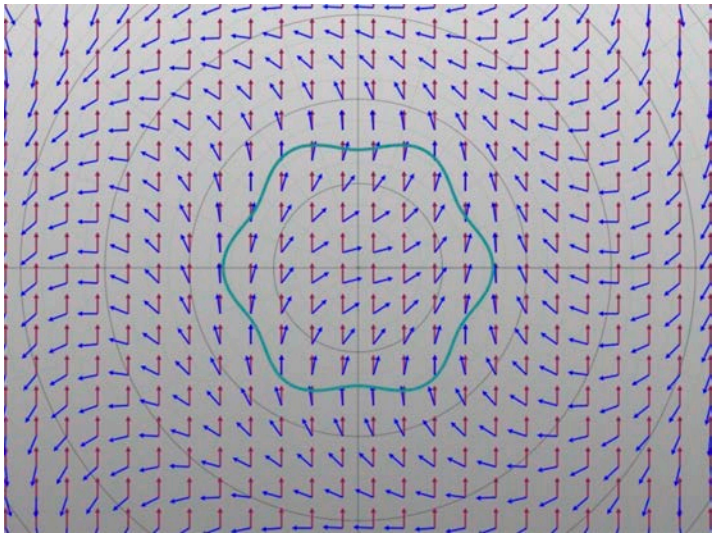


Figure II.3.37: *Super phase (B).* Phase of the superconducting condensate. This is again the trivial ground state but where the phase has been changed by a local gauge transformation, but the winding number is still zero.

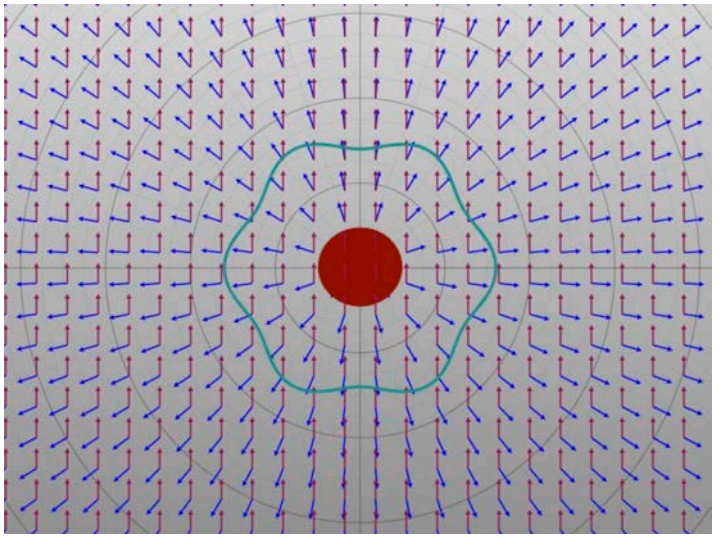


Figure II.3.38: *Super phase with flux (A).* Phase of the superconducting condensate with a magnetic flux tube in the center in the so-called radial gauge. The phase rotates by 2π after encircling the flux once along a closed curve like the green one in the figure. The winding number of this configuration equals $n = 1$.

start running so as to expel the magnetic field lines out the superconductor. This is the *Meissner effect*. In the type II superconductors, it is possible for flux lines to enter the medium, but only if the amount of flux satisfies a certain flux quantization condition. The situation is very similar to what we are discussing here: there is a superconducting ground state that corresponds to a condensate of pairs of electrons. These *Cooper pairs* have charge $2e$, and the medium has no electromagnetic field except for the filaments. The condensate is static and effectively described by a complex scalar field $\psi(\mathbf{x})$ that is doubly charged and carries an electromagnetic phase factor. To say that the pairs are condensed means that in that case the field acquires a constant non-zero magnitude, and because it describes the ground state it is called a *vacuum expectation* value. We write $|\psi(\mathbf{x})| = 1$ and ψ is described by a pure phase factor with angle $\beta(\mathbf{x})$. In Figure II.3.36 we have plotted the local phase β of the condensate in the ground state. The gauge field is in this case globally gauged to zero and the corresponding phase is trivial, $\beta(\mathbf{x}) = 0$. In Figure II.3.37 we have made a local ($\mathbf{x}$-dependent) gauge transformation which changes $\beta \to \beta + \Lambda(\mathbf{x})$. If we follow the phase along a closed curve like the green one, the phase will change forth and back, but the net change after returning to the initial point remains zero. We say that *winding number* of the configuration is $n = 0$. This winding number is not just gauge invariant. It a topological invariant, which means that it cannot be changed by *any* smooth transformation of the gauge potential or the phase $\beta(\mathbf{x})$. If we follow that phase around a magnetic flux line, the state should certainly return to the same value. It should be single valued because it is macroscopic state describing the condensate of Cooper pairs. We discussed this briefly in Chapter II.1 when discussing the *Josephson effect*. The upshot is that only fluxes are admitted that are 'invisible' for the medium, or the condensate. In other words, we want the induced phase factor to be equal one, which implies:

$$\exp(i\frac{2e}{\hbar c}\Phi) = 1 \quad \Rightarrow \quad 2e\Phi = 2\pi n\hbar c\,,$$

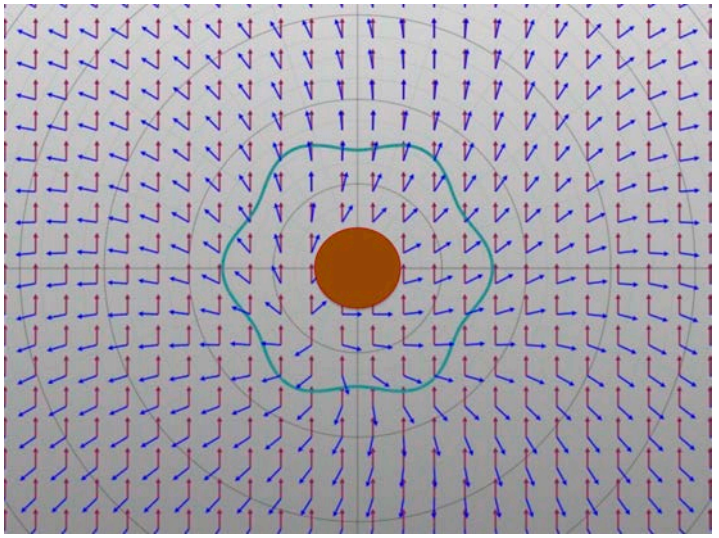


Figure II.3.39: *Super phase with flux (B).* Same physical situation as in the previous figure after a local gauge transformation, the winding number did not change.

which means that the flux is quantized according to:

$$\Phi = n\phi_o \;\; \text{with} \;\; \phi_0 = \frac{hc}{2e},$$

and ϕ_0 is called the fundamental flux quantum, which is expressed directly in fundamental constants. In the Figures II.3.38 and II.3.39 we show the phases of the condensate after a flux tube has entered. Moving along the green curve encircling the flux, the phase changes by 2π. This is clear in the radial gauge of the first figure but remains true after a gauge transformation has been applied.

This quantization rule is exactly what has been observed in type II superconductors. A flux tube has a negative surface energy and therefore an arbitrary flux likes to decay in individual minimally quantized filaments. These repel and therefore, if there is a strong magnetic field causing many tubes, these will form a lattice, a two-dimensional triangular crystal. If you keep turning up the magnetic field strength, the pairs will break up and therefore the superconductive state will break down at some critical value for the magnetic field.

The Berry phase

You may ask whether it is really possible to 'drag' a state vector along a closed loop like we described and whether the resulting phase change can be measured? The answer is affirmative. In this subsection we will discuss the *Berry phase* which is a substantial generalization of the Aharanov–Bohm phase, named after the British mathematical physicist Sir Michael Berry who discovered the possibility to measure holonomies in certain experimental setups with a well-chosen time or space dependent Hamiltonian.

The question is how to translate the rather abstract pictures of parallel transport into a suitable experimental setup. The idea behind Figure I.2.32 is clear: there is an 'agent' carrying the state vector, and by moving through space the frame changes and therefore the parallel transported vector appears to be rotated with respect to the initial local frame.

In the qubit or spin-one-half context you may think of the agent as an electron carrying a qubit (spin-one-half spinor) around. If we apply a magnetic field, the spin will align or anti-align with the external field as that minimizes the interaction energy. The ground state of the spin depends therefore on the orientation of the magnetic field. So to get the spin to move through its state space, we should move the electron in real space through an inhomogeneous magnetic field or we should fix its position and change the field. And by walking around along a closed loop in real space-time we may find the state of the spin is rotated by some phase angle. In other words, due to the inhomogeneous magnetic field, a closed loop in space-time gets mapped onto a smooth path in state space that is not necessarily closed.

In fact Berry took the approach where he looked at a time-dependent Hamiltonian $H(t)$. We have said that the time

evolution of a state is generated by the Hamiltonian. If the Hamiltonian is time=independent and the system is in an eigenstate of that Hamiltionian, then the time dependence is the time dependent phase factor $|\psi_n(t)\rangle = \exp(-iE_n(t-t_0)/\hbar)|\psi_n(t_0)$. The question is now what happens if the Hamiltonian becomes time dependent. You can think of the Hamiltonian having a set of parameters $\{c_i\}$. For example, if we consider the coupling of a spin to an external magnetic field, the parameters would correspond to choosing the direction and the strength of that external field. And the time-dependent Hamiltonian we are interested in would be one where we slowly vary these parameters: $H(t) = H(\{c_i(t)\})$. So the experiment is set up to see what happens if we make a round-trip through this parameter space or the space of Hamiltonians. The Hamiltonian moves between t_0 and t_f along a closed path in parameter space so that $H(t_0) = H(t_f)$. The choice of this path is of course made by the experimenter. In Figure II.3.40 we have depicted a time-dependent closed path (pink) through a two-dimensional coordinate space of Hamiltonians where $\mathbf{c}(t_0) = \mathbf{c}(t_f)$. The figure also shows the yellow path straight up, corresponding to the time independent Hamiltonian $H = H(t_0)$, leading to the aforementioned phase factor $\exp(-iE_n(t-t_0)/\hbar)$.

The expression for the phase factor. We assume that the Hamiltonian $H(t)$ has a time-dependent discrete spectrum:

$$H(t)|n(t)\rangle = E_n(t)|n(t)\rangle$$

If we now assume that we vary the Hamiltonian slowly so that the system smoothly (adiabatically) evolves in the state $|n(t)\rangle$, we can construct an approximate solution;

$$|\psi(t)\rangle = C_n(t)\exp\big(-\frac{i}{\hbar}\int_{t_0}^{t} E_n(t')\,|n(t)\rangle\,dt'\big),$$

and because $\psi(t)$ and $|n(t)\rangle$ are both normalized the coefficient $C_n(t)$ can only be a phase:

$$C_n(t) = \exp(i\gamma_n(t))$$

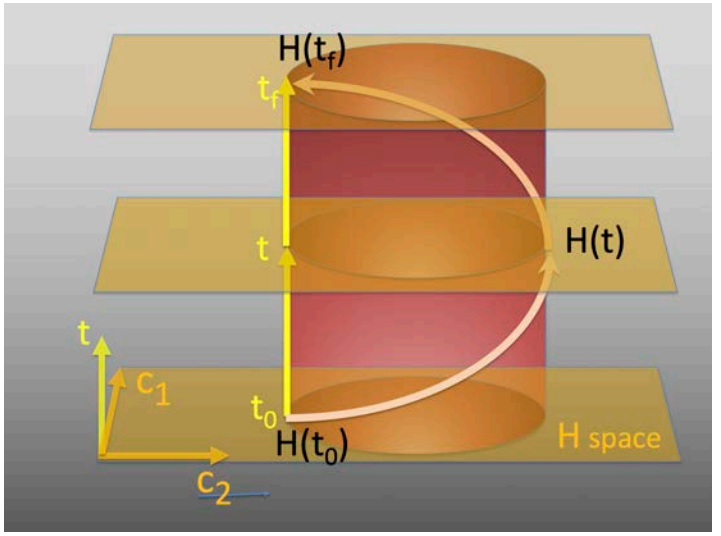


Figure II.3.40: *Berry phase.* A closed (circular) path in Hamiltonian space with coordinates $\mathbf{c} = (c_1, c_2)$. The system follows the pink curve in time such that $H(t_0) = H(t_f)$.

We can substitute this solution into the time-dependent Schrödinger equation,

$$i\hbar\frac{\partial}{\partial t}\psi(t) = H\psi(t)$$

to obtain an equation for the phase:

$$i\frac{\partial\gamma_n(t)}{\partial t}|n(t)\rangle = -\frac{\partial|n(t)\rangle}{\partial t}$$

which has the solution

$$\gamma_n(t) = i\int_{t_0}^{t}\langle n(t')|\frac{\partial|n(t')\rangle}{\partial t'}\,dt'$$

Berry connection and curvature. To give this phase a direct physical interpretation let us look at the integrand and ask what we mean by the state $|n(t)\rangle$. The time dependence is not the time dependence of n but rather of the state labeled by n. The time dependence all comes through the changing of the parameters $c_i(t)$. The appropriate notation is in fact to write $|n(t)\rangle = |n;\{c_i(t)\}\rangle =$

$|n;\mathbf{c}(t)\rangle$ where I combined the parameters into a vector, a position vector in parameter space. If I now take the time derivative of the state, I may rewrite that as follows:

$$\frac{\partial\,|n;\mathbf{c}(t)\rangle}{\partial t} = \boldsymbol{\nabla}_c\,|n;\mathbf{c}(t)\rangle\,\cdot\,\frac{\partial \mathbf{c}(t)}{\partial t}\,,$$

where the nabla operator is the vector of $\partial/\partial c_i$ derivatives. In other words the gradient operator acting on functions of the parameter vector.

This turns the time integral for the phase into a loop integral on parameter space over a connection (or pseudo gauge potential) $\mathbf{C}(c)$ named after Berry:

$$\gamma_n = \oint \langle n;\mathbf{c}|\boldsymbol{\nabla}_c\,|n;\mathbf{c}\rangle\cdot d\mathbf{c} \equiv \oint \mathbf{C}\cdot d\mathbf{c}$$

In other words, the phase factor using Stokes' theorem can be expressed as a surface integral of the corresponding Berry curvature $\mathbf{F} = \boldsymbol{\nabla}_c \times \mathbf{C}$:

$$\gamma_n = \oint \mathbf{C}\cdot d\mathbf{c} = \int \mathbf{F}\cdot d\mathbf{S}_c\,.$$

There is a striking analogy with the Aharonov–Bohm case, but it is also clear that the Berry analysis is much more general.

Spin coupled to an external magnetic field.

To be more concrete about such an experiment, imagine a closed path $\mathbf{c}(t)$ in time parametrized by a parameter $0 \le t \le 1$ with $\mathbf{c}(0) = \mathbf{c}(1)$. The system is an electron spin coupled to a slowly varying external magnetic field $\mathbf{B}(t)$, with a Hamiltonian

$$H(t) = \mathbf{B}(t)\cdot\sigma,$$

a hermitean 2×2 matrix acting on the two-component electron spin.

Let first ask what the space of Hamiltonians looks like, which is asking for a natural parametrization of all magnetic fields.

The field $\mathbf{B}(t)$ has some direction and some magnitude. As shown in Figure II.3.43 we choose spherical coordinates in $\mathbf{B}$ space. So the direction is parametrized by the angular coordinates θ and φ, while the magnitude is given by the radial coordinate. If we only change the direction of the external field, the space of possible Hamiltonians would just correspond to the radial magnetic fields on a spherical surface of constant radius. Note that this looks like the field surrounding a magnetic monopole as we have drawn in Figure I.1.29.

The starting point with the Berry phase experiment is to choose the time path that gets mapped onto some closed curve $\mathbf{c}(t)$ in the space of Hamiltonians, thus on the two--sphere in this case.

The adiabatic change or 'dragging' of the state amounts to parallel transporting a frame (of the tangent plane) along the curve, like we discussed in Chapter I.2 in the section on geometry.

As we will show shortly, the result for the acquired phase will depend on the solid angle that the path $H(t)$ has covered on the sphere.[4] This means that the Berry phase is a purely geometric phase (in fact a holonomy) which depends on the geometry of the space of Hamiltonians, but also on the probe (in this case a spinor).

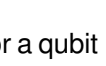
The idea is simple: at $t = 0$ we start at the North Pole with the Hamiltonian $H(0) = BZ$ and the energy eigenstates correspond $|\psi_n(0)\rangle = |\pm 1\rangle$. Next we start rotating the magnetic field and we assume that the initial eigen spinor just follows. In that sense it is fair to say that the Berry

[4] The path is oriented and the orientation decides whether to take the solid angle ω or $4\pi - \omega$, which with equation (II.3.4) amounts to $R_k(\theta) \to R_k(-\theta)$.

phase probes the Hamiltonian space but also the spin or qubit space which is a three-sphere S^3 as we know.

To work this out in more detail for an electron spin or a qubit for that matter we first look at how the rotations act on the spinors and then we find a convenient parametrization of the magnetic field space. ■

Probing the geometry of state space



To better understand what I mean by 'probing the state space' of a qubit I propose we return to the 'Barbie on a globe' representation of the qubit, as we introduced it in Chapter II.2 on page 286. What you see there is that we represent the qubit as a vector or rather spinor bundle over a two-sphere, where a particular qubit state corresponds to a unique tangent vector at some point on that sphere. And the X, Y and Z operators are generating the 'motions' of the Barbie in that space.

Let us first visualize the actions discussed above in the Figure II.2. We see the Barbie standing on the North Pole, say looking West corresponding to the state $-|1\rangle$. Acting with Z does not affect her at all, but acting with X moves her to the state $-|-1\rangle$ which is the mirror image of the initial state through the origin of the tangent space.

To probe the space in more detail we have to construct operators that move the Barbie around on the sphere and make her perform pirouettes. What we need are rotations generated around various axes, and these correspond to exponentials of X, Y and Z.

Rotation of qubits. As we will explain in more detail in the *Math Excursion* on page 635 of Part III on groups, this amounts to going from the *Lie algebra* of infinitesimal transformations to the corresponding *Lie group* of finite transformations. Here we need the explicit relation for any of the Pauli matrices $\sigma_k = X$, Y or Z that we introduced in

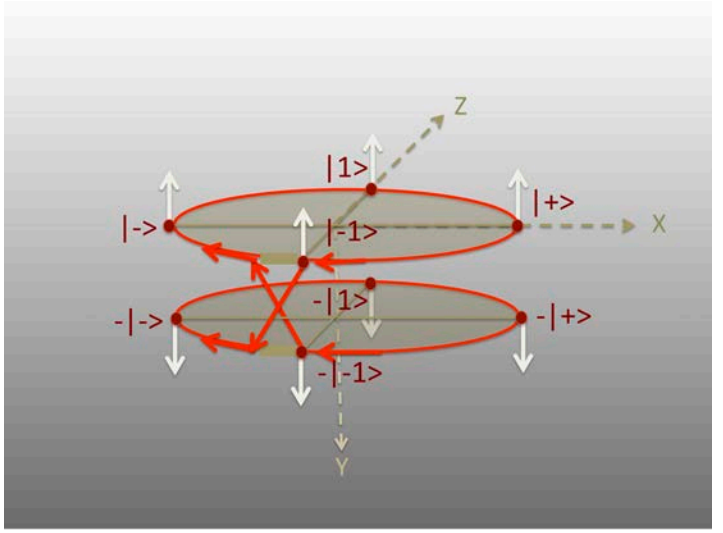


Figure II.3.41: *Effect of rotations around* Y*-axis on Barbie.* Rotating the Barbie state $|1\rangle$ around a large circle in the ZX-plane by angles $\theta = \pm n\pi/2$. The rotation angle in this representation equals θ. As she only passes through real states, the overall phase β, denoted by the white arrow, is either 0 or π. Rotating over an angle 2π any state goes to minus itself.

equation (A.34) of the chapter on *Math Excursions*:

$$R_k(\theta) \equiv \exp\left(i\frac{\theta}{2}\sigma_k\right) = \mathbf{1}\cos\frac{\theta}{2} + i\sigma_k \sin\frac{\theta}{2}\,, \qquad \text{(II.3.4)}$$

which should be compared with its one-dimensional analogue, the Euler formula (A.28). At this point we recall some important observations we made before.

1. Since the spinor or qubit is a two-dimensional complex vector, the rotations are relatively simple two-by-two unitary matrices which can be given explicitly as you see.

2. These complex rotations form the group $SU(2)$.

3. The formula for $R_k(\theta)$ represents a rotation about the k-axis over an angle θ. That means to say that acting on an ordinary three-dimensional vector like $\mathbf{B}$, it rotates over an angle θ in real space. That is, under

a rotation $R_k(\theta)$, , the Hamiltonian will rotate as:

$$H \to H' = \mathbf{B} \cdot \sigma' = \mathbf{B} \cdot R_k \sigma R_k^{-1} \qquad \text{(II.3.5)}$$

4. However, on the two-component complex state vector of a qubit the rotation acts like

   $$|\psi\rangle \to |\psi\rangle' = R_k\,|\psi\rangle.$$

   Note that it 'rotates' only over half that angle because of the factor $\theta/2$ in the formula (II.3.4).

5. This factor one-half has dramatic consequences. For example a rotation by $\theta = 2\pi$ around any axis produces in (II.3.4) just minus the unit matrix! So, under such a transformation the qubit state always goes to minus itself. One has to rotate by 4π before one gets back to the unit matrix. This is indeed a defining property for a *spinor*, to be contrasted with rotating an ordinary vector about an angle 2π, which always gives the same vector back. This minus sign for a spinor has a deep physical significance for particles carrying half-integer spin as we will explain later. It is one of those minus signs that does matter a great deal.

6. Note that under the rotations the norm of the state is preserved

   $$\langle\psi'|\psi'\rangle = \langle\psi|\psi\rangle\,,$$

   and this is what we expect as we are moving over a sphere, because by taking the complex conjugate vector the transformation is going to its hermitian conjugate, which means changing $i \to -i$ or $\theta \to -\theta$, what amounts to the same thing. This means that the conjugate rotates by the opposite amount, so that the net effect of the rotation on the inner product of vector with itself (or any other vector) always cancels.

To familiarize ourselves a bit with these rotations, let us first restrict our attention to real qubit state vectors as introduced in Figure II.1.7. These states form a real circle. The operators Z and X are qubit observables which have real eigenvectors. For Z those are $\pm|1\rangle$ and $\pm|-1\rangle$ respectively with eigenvalues $+1$ and -1. Similarly for X we have $\pm|+\rangle$ and $\pm|-\rangle$ also with eigenvalues $+1$ and -1 respectively. We may ask which operator would move you around in that subspace of real states, on that circle. Such moves correspond to rotations about the Y-axis, generated by Y and indeed, a rotation by an angle θ yields the real matrix:

$$R_y(\theta) = \cos\frac{\theta}{2} + iY\sin\frac{\theta}{2} = \begin{pmatrix} \cos\frac{\theta}{2} & \sin\frac{\theta}{2} \\ -\sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{pmatrix}, \qquad \text{(II.3.6)}$$

which indeed corresponds to a rotation of the qubit state vector over an angle $\theta/2$. In other words, rotations about the Y in the $(|-1\rangle, |1\rangle)$ plane move a state around the circle.

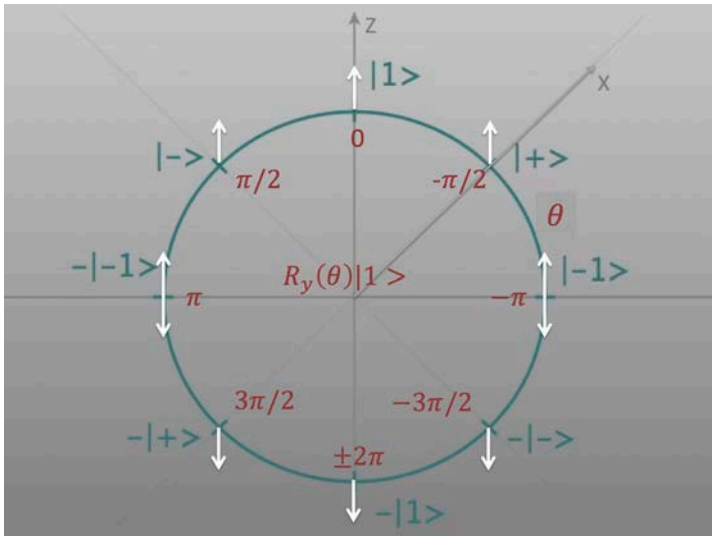


Figure II.3.42: *Effect of rotations on the real circle of qubit states.* Same situation as previous figure, but now we have plotted the states at half the angle θ from $-2\pi < \theta < 2\pi$. In the upper half-of the circle the overall phase β is for real states 0 and in the lower half it is π.

So let us find out what the formula yields for rotations over multiples of π acting on the $|+1\rangle$ state, and then visualize

the results in the two different representations of the qubit space corresponding to (i) the 'Barbie on the globe' picture, and (ii) the real circle of Figure II.1.7.

Using the formula (II.3.6) we obtain the following values for some ($-2\pi \leq \theta \leq 2\pi$) rotations. For a transformation by $-\pi$ we find:

$$R_y(-\pi)|1\rangle \leftrightarrow \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \leftrightarrow |-1\rangle\,,$$

and for the others:

$$R_y(\pm\frac{\pi}{2})\,|1\rangle = |\mp\rangle, \quad R_y(\mp\pi)\,|1\rangle = \mp|-1\rangle\,,$$
$$R_y(\pm\frac{3\pi}{2})\,|1\rangle = -|\pm\rangle, \quad R_y(\pm 2\pi)\,|1\rangle = -|1\rangle\,.$$

If we carry a spinor along a large circle over an angle of 2π we obtain from (II.3.6), just a (phase)factor minus one. We have illustrated the sequence of values just calculated in Figure II.3.41 which should be compared with the 'Barbie on the globe' figure on page 286. The rotations for increasing values of θ correspond to the Barbie moving by the same angle over the globe, anti-clockwise in the vertical plane. The states remain real for all θ and the only phase change that may occur is that it jumps from 1 to -1 or the other way around. This corresponds to a jump in the phase angle of $\beta = \pm\pi$ depicted by the white arrows either pointing up or down in the figures.

In Figure II.3.42, the same sequence is represented in the standard qubit decomposition that we introduced in Figure II.1.7, and we see indeed the phase jumping at odd-multiples of $\theta = \pm\pi$.

You may think of this as a *holonomy* effect, referring to the concept we introduced in Chapter I.2 while discussing parallel transport of vectors through curved space, which is exactly what we are doing here. If the Barbie parallel transports her spinor, it may pick up a phase factor equal minus one. So if she starts walking along a big circle on the sphere looking straight ahead, she will looking straight back upon returning. What the Figures II.3.41 and II.3.42 show is that the Barbie at $\theta = \pm\pi$ suddenly turned her head by 180^o ..

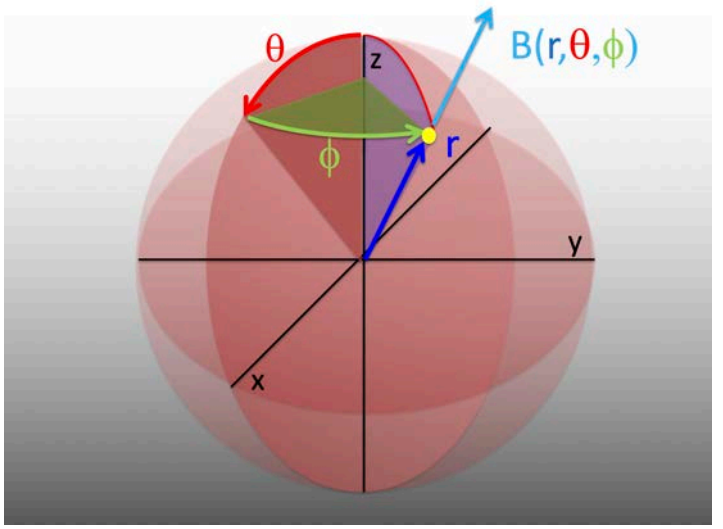


Figure II.3.43: *Radial magnetic field.* The Hamiltonian landscape.

Magnetic field space. We choose that initially the field is in the positive z-direction $\mathbf{B}(0) = B\hat{\mathbf{z}}$ and the spin to be in the aligned up $|1\rangle$ state, so, in the $n = 1$ energy eigenstate. We change the direction of the field slowly so that the spin stays aligned with the varying external field provided the changes are slow.

From the figure we learn what the x, y and z components of $\mathbf{B}$ are in terms of the angular variables:

$$B_x = B\sin\theta\cos\varphi \quad B_y = B\sin\theta\sin\varphi \; B_z = B\cos\theta$$

And thus the Hamiltonian of equation (II.3.5) corresponding to a point on the sphere i (we set $B = |\mathbf{B}| = 1$) looks in matrix form like:

$$H(\mathbf{c}) = H(\theta,\varphi) = \begin{pmatrix} \cos\theta & e^{-i\varphi}\sin\theta \\ e^{i\varphi}\sin\theta & -\cos\theta \end{pmatrix}.$$

The two eigenstates with eigenvalues plus and minus one

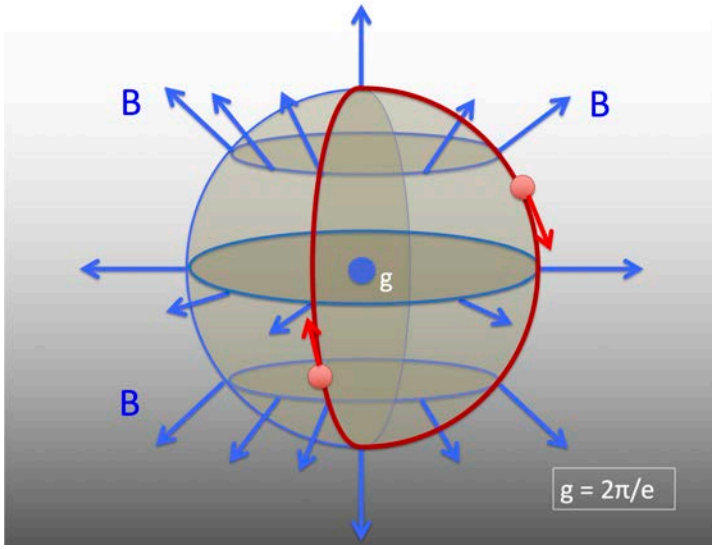


Figure II.3.44: *Magnetic field space.* The space of a magnetic field of constant magnitude can be represented as the field on a sphere around a magnetic monopole. Adiabatic transport of a spin-1/2 particle moving along a closed path (in red) in the radial magnetic field of a pole of strength $g = 2\pi\hbar/e$ centered at the origin (blue).

correspond to the spinors:

$$|1;\mathbf{c}\rangle = \begin{pmatrix} \cos\frac{\theta}{2} \\ e^{i\varphi}\sin\frac{\theta}{2} \end{pmatrix} \quad \text{and} \quad |-1;\mathbf{c}\rangle = \begin{pmatrix} -e^{-i\varphi}\sin\frac{\theta}{2} \\ \cos\frac{\theta}{2} \end{pmatrix}$$

The adiabatic process. The process of adiabatically moving over the sphere corresponds to making rotations around the proper axis and angles. So for example to move from the North Pole to the point (r,θ,φ) one may apply the transformation(s):

$$R_B = R_z(-\varphi)R_y(-\theta)R_z(\varphi) = \begin{pmatrix} \cos\frac{\theta}{2} & -e^{-i\varphi}\sin\frac{\theta}{2} \\ e^{i\varphi}\sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{pmatrix}.$$

One checks that:

$$H(\theta,\varphi) = R_B\, Z\, R_B^{-1} \quad \text{and} \quad |\pm 1;\mathbf{c}\rangle = R_b\,|\pm 1\rangle\,,$$

as it should be. ■ ■

The Berry connection. 

Let now calculate the Berry connection which involves applying the ∇_c operator, but the c coordinates are just the ordinary three-dimensional spherical coordinates where the (angular) components are given by $\nabla_\theta = \partial/\partial\theta$ and $\nabla_\phi = \sin^{-1}\theta\partial/\partial\varphi$. The Berry connection is then:

$$\mathbf{C}(\theta,\varphi) \equiv \langle n;\mathbf{c}|\nabla_c\,|n;\mathbf{c}\rangle = \frac{1-\cos\theta}{2\sin\theta}\,\hat{\varphi}\,.$$

This connection is exactly the gauge potential written down by Dirac in his famous 1931 monopole paper, and indeed its curl give the field of a magnetic monopole with magnetic charge $eg/\hbar c = 2\pi$. The total magnetic flux through the sphere is 2π, which is half the solid angle of the total sphere being 4π. And thus is the resulting phase after closing the loop equal to the magnetic flux going through the loop. It is nice to see how nice this subject of the Berry phase connects with matters that we discussed in early chapters of Volume I.

Some explicit examples. Let us now consider some specific paths and see how this works. In the first example we start at the North Pole meaning to say that $H = Z$ and $|\psi(t=0)\rangle = |1\rangle$. Then we parallel transport vector along a geodesic generated by rotating around X-axis over an angle $\theta = -\pi$ and bring it back along a geodesic generated by rotating around the Y-axis by $\theta = \pi$. The path corresponds to the red two-angle indicated in Figure II.3.44. This means that the Hamiltonian between $t = 0$ and $t = \frac{1}{2}$ smoothly rotates in the YZ-plane from $Z = H(0)$ to $Y = H(\frac{1}{2})$. From formula (II.3.4) we see that:

$$R_k(\pm\pi) = \pm i\sigma_k\,.$$

So the overall (unitary) transformation of the state vector corresponds to:

$$U = iY(-iX) = -iZ\,.$$

So the net effect on the state $|1\rangle$ after coming back home is that it is rotated by an angle $\theta = -\pi/2$ around the z-axis. So the loop integral would give a magnetic flux of

$\pi/2$ which is 1/4 of the total flux of 2π, which in turn is consistent with the fact that the loop covered 1/4 of the total solid angle of the sphere.

So what is the interference effect on the probabilities measured, if we start with $|\psi_1\rangle$ and end up at $|\psi_2\rangle = -iZ|\psi_1\rangle$? The expression is given by the following equation for any outcome of a measurement:

$$I(a_i) = \gamma(\langle\psi_1|P_i^A|\psi_2\rangle + \langle\psi_2|P_i^A|\psi_1\rangle)$$

The outcome is the expectation value of an expression involving U and P_i^A:

$$I(a_i) = \gamma\langle\psi_1|\,(P_i^A U + U^\dagger P_i^A)\,|\psi_1\rangle\,.$$

We obtain that in the case at hand by choosing $P = P_\pm^Z$ the result is zero for all $|\psi_1\rangle$. However for $P_\pm^Y$ we find $\langle\psi_1|Y|\psi_1\rangle$ which of course may or may not be observable dependent on the choice of the initial state.

Another example would be as sketched in Figure I.2.32. There we have three successive rotations by $\pi/2$.

$$U = R_y(\frac{\pi}{2})R_z(\frac{\pi}{2})R_x(\frac{\pi}{2}) = \frac{1}{\sqrt{2}}(1 + iZ)\,,$$

where we used that

$$R_k(\frac{\pi}{2}) = \frac{1}{\sqrt{2}}(1 + i\sigma_k)\,.$$

This generates interference in more situations than the previous case, and applying it to $|\psi_1\rangle = |1\rangle$ we get:

$$U\,|1\rangle = \frac{1}{\sqrt{2}}(1 + iZ)\,|1\rangle = \frac{1}{\sqrt{2}}(1 + i)|1> = e^{i\pi/4}\,,$$

again this is consistent with 1/8 of the total flux. ■ ■

Quantum tunnelling: magic moves

In this chapter we have considered the consequences of the quantessential particle-wave duality in typical wave type phenomena like reflection, refraction, diffraction and interference. In this section we turn to the aspect of transmission, notably the effect of quantum tunneling, which is another stunning instance where quantum theory overrides a classical veto. In the tunneling process we should think of particles that can move through, or jump over a potential wall. This happens for example in the spontaneous decay of bound systems, and has a great application in scanning tunneling microscopy (STM). Such processes are strictly forbidden by classical physics but have finite although small (meaning exponentially small) probabilities to occur in the quantum situation. It can be looked upon as a consequence of the quantum fluctuations in the system that 'follow' from the uncertainty relations.

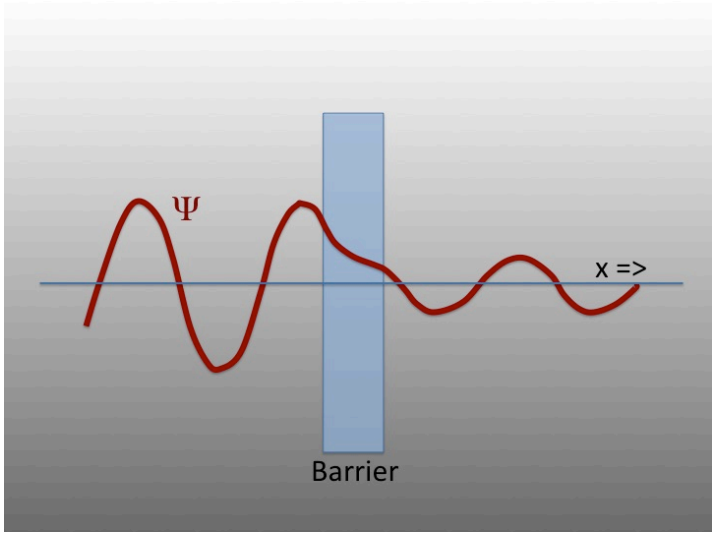


Figure II.3.45: *Quantum tunneling.* The lowest energy state of a particle in the presence of a potential wall shows that the quantum particle is most probably found on the left-hand side, but still has a small probability to be on the right-hand side. The wavefunction decays exponentially in the wall but still has a non-vanishing value when it arrives at the other side.

Let us put a quantum particle in a bowl corresponding to a potential energy landscape as given in Figure II.3.45. Imagine the particle sitting in the origin at the bottom of

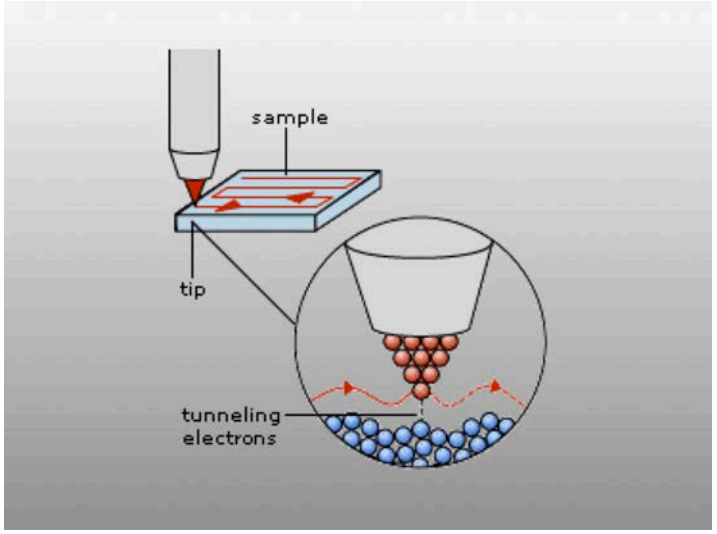


Figure II.3.46: *The scanning tunneling microscope.* Fixing the tunneling current, fixes the distance between the tip and the surface to be scanned. (Source: Flickr.)

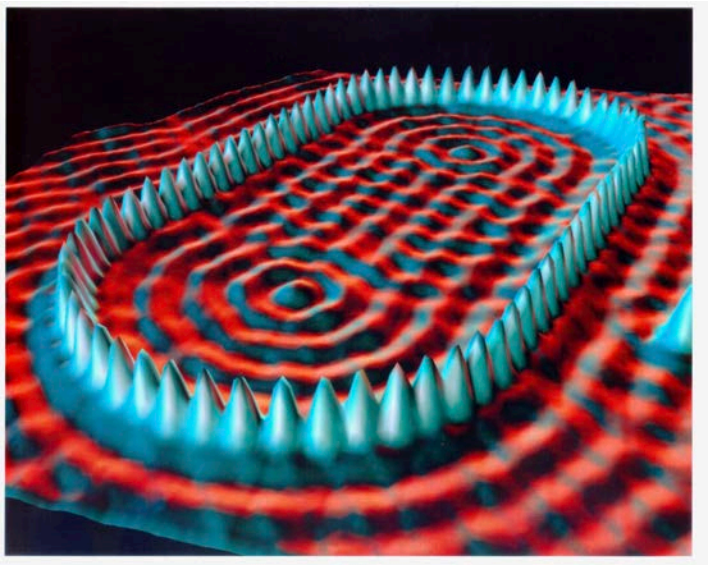


Figure II.3.47: *STM surface imaging.* Image of scanning tunneling microscope of a coral of atoms deposited on a surface. (Source: IBM.)

the bowl. Then clearly, if we do not add enough energy to overcome the height of the bowl, it will sit there forever, since from a classical point of view it is a stable situation. However, in the quantum world the problem is different; the lowest energy solution for the wavefunction is sketched in red and the important point to observe is that it is non-zero *outside* the bowl. In other words, if we square the wavefunction we get a large probability to find the particle where we expect it, but there is a non-vanishing probability of finding the particle outside. There is a small probability for the particle to 'jump' the wall to the outside world, where we might observe it. It jumps a wall of any height as long as it is thin enough.

In more physical terms you may think of a situation where a particle is bound (and thus sitting in some potential well), but if the well corresponds to a local minimum, then there is a (low) probability that the particle will tunnel out of the well, meaning that the system decays and emits the particle. This is for example what happens with nuclear α decay, certain nuclei will spontaneously emit an α particle which is in fact just a ^4He nucleus consisting of two protons and two neutrons. It is this tunneling phenomenon that explains the extremely – exponentially – small probabilities that are reflected in the extremely long half-life times of certain nuclei. Long means that the process takes place with a much lower frequency than the natural frequency f that occurs in the state corresponding with the energy E of the state, with $E = hf$.

A similar situation occurs if one sends a particle to a potential barrier (a wall) then classical physics may predict some energy and momentum transfer during the impact by which the particle is stopped or may be reflected, but what we never have is that the particle would have a finite chance to of moving through the wall (without destroying it). And this is exactly what happens in the quantum case, where one finds a definite probability of ending up on the other side as long as the wall has a finite size. The reflection and transition probabilities can be calculated and of course add up to one.

We discussed a realization of tunneling currents in chapter II.1 for the Josephson junction. Another important application is the scanning tunneling microscope (STM). The working is schematically depicted in Figure II.3.46. By driving a constant small current through the tip to the surface one wants to study, the tip will then precisely scan the surface, all the way down to atomic scales. The tip will never touch the surface and the 'wall' is provided by the thin insulating layer of air between the tip and the surface. The images taken by the microscope of the surface localizes the presence of isolated atoms or molecules on the surface. A nice example is given in Figure II.3.47. The STM scans the contour of the charge density profile on the surface. People can be stopped by virtual walls, but walking through a real wall is quite something else, and that is what quantum particles apparently can do.

Further reading on interference:

- *QED:*
  The Strange Theory of Light and Matter
  Richard P Feynman Antony Zee
  (revised version)
  Princeton University Press (2014)
- *Quantum Interference and Coherence: Theory and Experiments*
  Zbigniew Ficek and Stuart Swain
  (Springer) (2005)

Chapter II.4

Teleportation and computation

Entanglement and teleportation

The Einstein–Podolsky–Rosen paradox

In 1935 Albert Einstein, Boris Podolsky and Nathan Rosen, confronted quantum physics with a profound objection concerning the quantessential property of entanglement. This led to a fierce debate between Bohr and Einstein closely followed by Schrödinger. In those days the problem was presented as a *gedanken* experiment involving a pair of spins or qubits which are entangled but widely separated in space. One may think of a spin-less particle at rest (a π_0 particle for example) decaying into two photons, because of momentum conservation both particles will fly off back to back and because of spin conservation the polarizations of the two photons have to be opposite. This means that without interactions the particles could separate and travel a long way, and we could imagine that one might arrive in New York and the other in Tokyo where Alice and Bob will make polarization measurements. The polarization state of the entangled pair is given by:

$$|\psi_{\rm NT}\rangle = \frac{1}{\sqrt{2}}(|1,-1\rangle - |-1,1\rangle)\,, \qquad \text{(II.4.1)}$$

where the first entry refers to the NY particle and the second to its Tokyo counterpart, and we for convenience have assumed the particles to be polarized in z-direction. Now

Figure II.4.1: *The Myth of Depth*, a 1984 painting by Mark Tensey. It makes you think of unusual, if not magical, ways information may propagate. It is the 'Spooky action at a distance,' Einstein was so worried about.(Source: ANP / Mark Garlick / Science Photo Library)

Alice in New York decides to make a polarization measurement. Let us suppose that she chooses to do this along the x-axis, and let us also suppose that she finds a value $+1$. Then we know that the first spin is projected on the $|+\rangle$ state. But as the spins are opposite it follows that instantaneously the spin of the particle in Tokyo must have changed to the $|-\rangle$ state. That this indeed has to be the case follows from the fact that we could have written the

initial state also in the form

$$|\psi_{\rm NT}\rangle = \frac{1}{\sqrt{2}}(|-,+\rangle - |+,-\rangle)\,,$$

and Alice's projects on the first term as we discussed in the previous section, so after Alice's measurement we have $\psi_{\rm NT} \Rightarrow |+,-\rangle$. If Bob also decides to measure along the x-axis, then he will obtain the value -1. It is clear that the probabilities for measurement outcomes can be precisely calculated for all possible independent choices that Alice and Bob could make.

There is a lot at stake in this proposed experiment and in the early days was it was too hard to perform. If the calculated and observed distributions would not match, then quantum theory would be in deep trouble, not to say falsified! As we will discuss later, starting in the 1980s, such experiments became feasible, and in fact unambiguously confirmed the quantum predictions.

Quantum key distribution. The above observations allow for a quite simple protocol to securely share a digital key, called the BB84 protocol, which was invented by Gilles Brassard and Charles Bennett in 1984, opened a research field in quantum informatics called quantum cryptography. Their protocol benefits from the fundamental principles of quantum mechanics and enables secure communication between parties. Nowadays, their protocol is commercially available and forms the core of many other protocols on quantum cryptography and quantum information in general. Brassard and Bennett shared the prestigious Breakthrough Prize in Fundamental Physics 2023 with David Deutsch and Peter Shor.[1] The Shor quantum algorithm for prime factorization will be discussed in the next section on quantum computation. The protocol is illustrated in Figure II.4.4. Alice and Bob take a large sequence of measurements on (in this case) parallel polarized entangled pairs and make a list of their sequence of polarizer settings and their outcomes. Afterward they may exchange the sequences of their polarization settings. If they now select the outcomes for the pairs where the setting was identical, then the outcomes must be the same, therefore this restricted sequence represents a shared digital code quantum computation as may be verified from the figure.

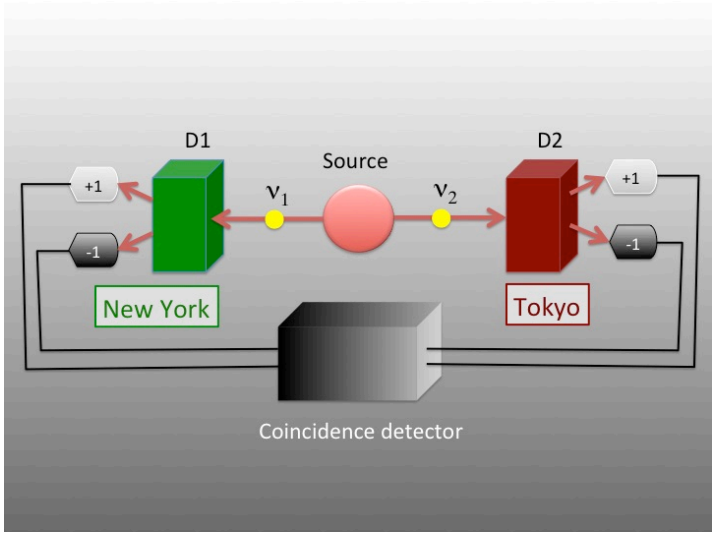


Figure II.4.2: *The Einstein–Podolsky–Rosen (EPR) paradox.* Two particles are created in a polarization entangled state, and a measurement outcome on the left particles completely determines the probabilities for the measurement outcomes on the right in any frame. The coincidence detector is there to make sure that measurements on members of the same pair are compared.

If one imagines an eavesdropper Eve somewhere measuring one of the photons, she cannot copy it and resend it. This means that the observed code that Alice and Bob observe will no longer coincide. So they can check whether their communication channel is secure. Clearly Eve cannot extract any key from her observations.

Is causality violated? Einstein's first worry was that this instantaneous non-local consequence of the act of mea-

[1] The Breakthrough Prize in Fundamental Physics is one of the largest prizes in science – both qua money and prestige – and was founded in 2012 by Yuri Milnor.

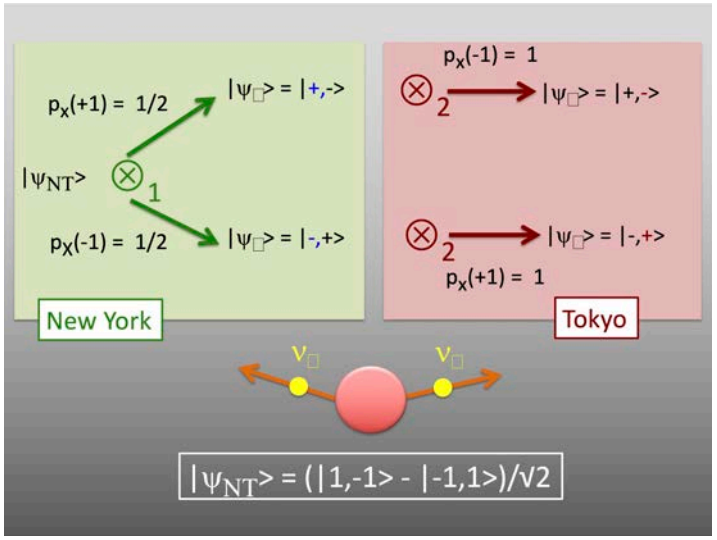


Figure II.4.3: *EPR schematic.* The measurement scheme for a particularly simple choice of measurements in the EPR experiment. The pair is created in the state $|\psi\rangle$ with opposite polarization in the z-frame. Alice in New York measures in the x-frame, so she finds outcome ± 1 with equal probability $p_x(\pm 1) = 1/2$. If Bob in Tokyo subsequently also measures in the x-frame, his outcome, according to quantum theory is completely fixed.

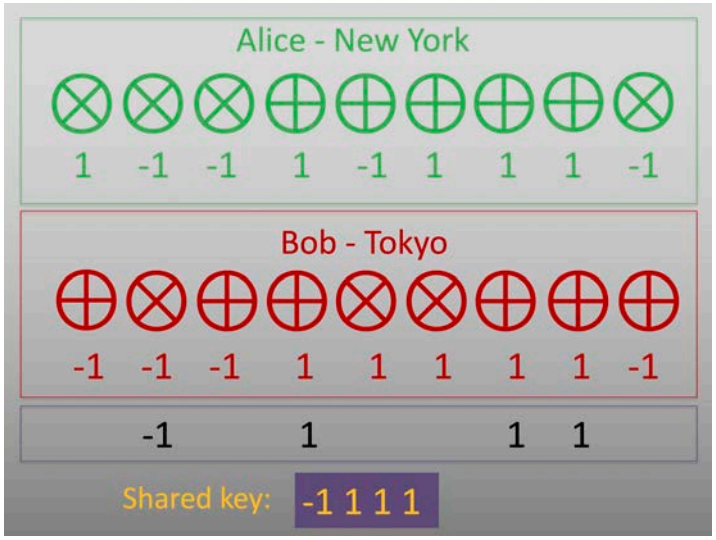


Figure II.4.4: *Quantum key sharing.* Using a sequence of parallel entangled photons for key distribution through the BB84 protocol. On top in green is the sequence of polarizer choices that Alice made and in the second line her measurement outcomes. In the red box we give the sequence of Bob and his outcomes. What we know for sure is that when the members of a pair are measured in the same polarization frame, the outcomes should be identical. And indeed, if we cross out the measurements where the frames are different, we are left with two identical sequences. If this happens not to be the case, Alice and Bob know that an eavesdropper is active somewhere.

surement of one of the particles of the entangled pair, would violate causality. Some information about Alice's measurement outcome appears to have been transmitted instantaneously to Tokyo, which means that it had to travel with a velocity exceeding the speed of light. And that is a no-go in Einstein's relativity!

So, the first task is to actually prove that the correlations between the measurement outcomes would necessarily require the transfer of information faster than light. If so, this would mean that such pairs could be used to transmit information faster than light, which in turn would imply the breakdown of special relativity in particular and of our cherished notion of causality in general.

The question should be: what can Bob learn from Alice making a measurement? As a matter of fact, the answer is: *nothing at all*, at least as long as he doesn't know what the polarization axis is that she has chosen for her measurement, and what the outcome of her measurement was. But she can only inform him about that by conventional means using subluminal velocity media like email or Facebook. So this form of information sharing does not violate causality.

Hidden variables and local realism. The proposition of the EPR trio was that quantum theory, which clearly was in accordance with all available observations, was maybe not really wrong but at least incomplete. The paradox furthermore implied that once completed the theory would not need these 'spooky' instantaneous non-local kind of inter-

actions. Any physically sound theory should obey the principle of *local realism*. Local realism maintains that each of the particles is always in a definite polarization state all along, but it just happens to be so that we don't know which state that is. The state is always completely determined but we don't know how it is fixed. Maintaining local realism would be possible if you say that the highly correlated nature of the outcomes could be a manifestation of ordinary statistics caused by the existence of certain *hidden variables*, which would cause such correlations. The need for that strange, non-local, instantaneous 'action at a distance' could be avoided if one knew these hidden variables and would measure them. In other words, Einstein was not arguing about the predictions of quantum theory *per se*, but the proposed probabilistic formalism would only be part of the story – a kind of effective description of nature, and not a fundamental ingredient of the resulting complete theory. His proposal would turn the fundamental indeterminacy of quantum theory merely into a lack of knowledge about the set of state variables. A fundamentally *undetermined* state would just become a fully determined, but *unknown* state.

This line of reasoning caused a rather deep controversy about the measurement problem and the interpretation of quantum theory. Because the tremendous successes of quantum theory continued to unfold, this Einstein–Bohr debate lingered on somewhat in the margins as a kind of pastime for philosophers of science, until in 1964 John Steward Bell, a British physicist working at the CERN accelerator center in Geneva, made the groundbreaking discovery that there are situations where quantum theory would directly contradict the local realist predictions. Bell turned Local Realism into a falsifiable hypothesis! The question was to set up a true EPR experiment and precisely measure the correlations between the measurement outcomes for the entangled pairs. Bell's proposal moved the question out of the realm of abstract epistemology into that of experimental physics. This deep question allowed for a definite answer. This is the subject of the next section.

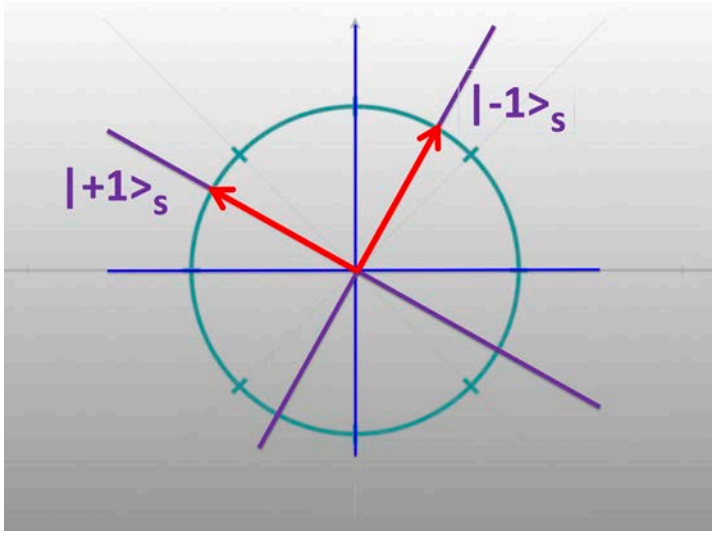
The Bell inequalities

The discussion of Bell is about the EPR pairs and the measurements illustrated in Figure II.4.5. The question is indeed whether a hidden variable theory could ever account for the data as predicted by quantum theory. Is there a deterministic scheme which respects local realism that perfectly mimics the quantum theory and the measurements on entangled states? The difficulty is in some sense to produce the extremely strong instant correlations between measurement outcomes that quantum mechanics allows, even if the particles are far apart.

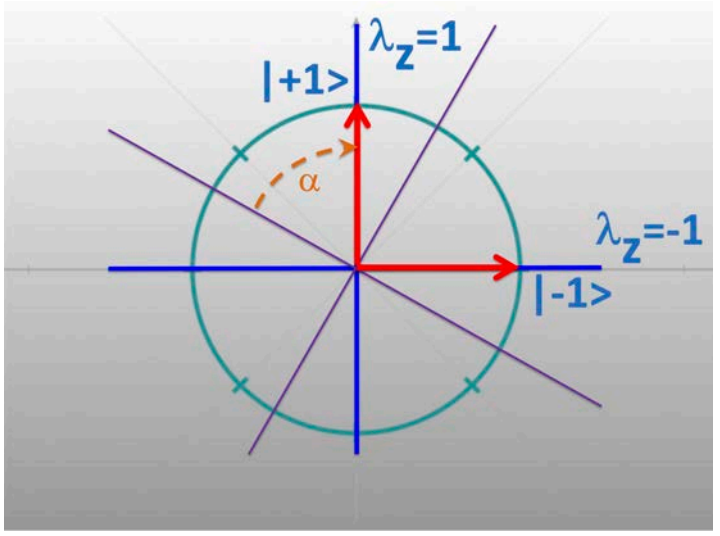
The correlator. John Bell devised an experimental test exactly based on these correlations. To stay in the language of the previous section, Bell proposed to consider the average of the product of measurement outcomes of Alice and Bob $\mathrm{P}(\mathfrak{a},\mathfrak{b})$ where $\mathfrak{a}$ and $\mathfrak{b}$ are the (real) frames of Alice and Bob as depicted in Figure II.4.5(d). If we imagine that they both choose the same polarization, one finds for example that:

$$\mathrm{P}(\mathfrak{a},\mathfrak{a}) = -1 \text{ and } \mathrm{P}(\mathfrak{a},-\mathfrak{a}) = 1\,, \qquad \text{(II.4.2)}$$

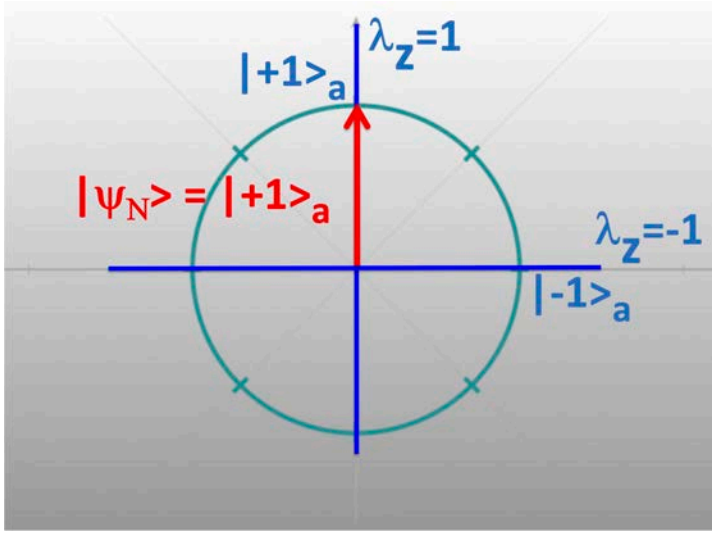
because if they have the same frame the measurement outcomes will be opposite and the product becomes minus one. If the polarizers $\mathfrak{a}$ and $\mathfrak{b}$ are in the same direction but oriented oppositely, they both measure $+1$ and thus the correlator is plus one. Now it is clear that to calculate the correlator $\mathrm{P}(\mathfrak{a},\mathfrak{b})$ in general for the quantum case, we just have to look at the figure, where we learn that if the angle between the frames of Alice and Bob is β and Alice measures $+1$ then Bob measures $+1$ with probability $\mathrm{p}_{\mathfrak{b}}(+1) = \sin^2\beta$. This is consistent with equation (II.4.2), because $\mathrm{P}(\mathfrak{a},\mathfrak{a}) = -\cos 0 = -1$ and $\mathrm{P}(\mathfrak{a},-\mathfrak{a}) = -\cos\pi = 1$, and similarly $\mathrm{p}_{\mathfrak{b}}(-1) = \sin^2\beta$. And if Alice measures -1 then also the probabilities $\mathrm{p}_{\mathfrak{b}}(\pm 1)$ get interchanged. From these considerations one obtains the



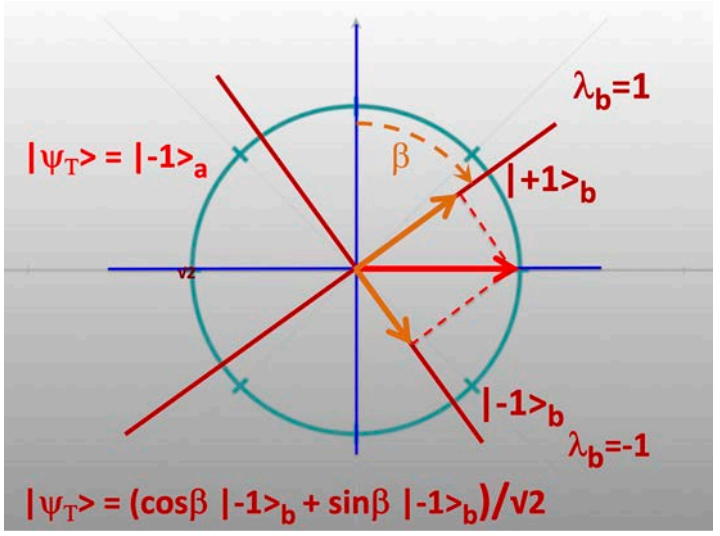
(a) The electron-positron pair is produced in some frame σ_s in the antisymmetric entangled state $|\psi_{NT}\rangle = \frac{1}{\sqrt{2}}(|1,-1\rangle_s - |-1,1\rangle_s)$, which is represented by the double arrow.



(b) The electron-positron pair state in the frame $\sigma_a = \sigma_z$ of Alice in New York. The antisymmetric form is preserved under rotations, and we just replace the subscript s with a.



(c) The spin measurement in σ_a frame of Alice in New York. She measures the eigenvalue $\lambda_a = +1$, and projects the New York component on the $|+1\rangle$ state.



(d) After Alice's measurement the Tokyo component collapses to $|\psi_T\rangle = |-1\rangle_a$, from which the probabilities for the measurement outcomes in the σ_b of Bob follow.

Figure II.4.5: *The Einstein–Podolsky–Rosen paradox.* (a) A neutral particle decays into an entangled electron-positron pair; these travel in opposite directions to New York and Tokyo and have oppositely polarized spins in some frame. Alice and Bob make subsequently measurements in frames they may choose independently and each will measure an outcome ± 1 . The sequence of subfigures explains that the final probability for Bob is $\sin^2\beta$ to find $+1$, and $\cos^2\beta$ to find -1 . These probabilities depend on Alice's choice and are instantly fixed after Alice has made her measurement.

following formula for $P(a,b)$:

$$\begin{aligned} P(a,b) &= \frac{1}{2}(1(\sin^2\beta - \cos^2\beta) - 1(\cos^2\beta - \sin^2\beta)) \\ &= \sin^2\beta - \cos^2\beta = -\cos 2\beta\,. \end{aligned} \tag{II.4.3}$$

Introducing hidden variables. To describe the measurements in hidden variable theory we can introduce two functions $A(a,\lambda)$ and $B(b,\lambda)$ representing the measurement outcomes of Alice and Bob respectively, which are strictly local in the sense that they only depend on their own measurement frame, and now also on a hidden variable λ. A value for this variable is typically set at the moment when the particles are produced and that value stays fixed for both in the absence of interactions. Given λ and a choice of frame a, the outcome $A(a,\lambda)$ is fixed. The question is whether there exist such functions that reproduce the quantum results of equation (II.4.3). Here Bell brilliantly rephrased the question. Instead of just trying to directly prove or disprove the existence, he derived a condition (in fact a bound or inequality), which any hidden variable theory would have to satisfy under quite general assumptions, and subsequently showed that quantum theory allows for ample situations where this condition would be violated. Answering the question was now reduced to performing certain experiments and seeing whether the results would violate the inequality or not. If they do not, hidden variables would be a viable option, but if they do, that would be the demise of the theory of hidden variables and local realism!

The Bell inequality. Let us first agree that A and B can only equal ± 1, because they are measurement outcomes. The only thing we assume about λ is that it can take certain values with a probability $w(\lambda)$, where we have to require that $w(\lambda) \geq 0$ and $\Sigma_\lambda w(\lambda) = 1$. The classical 'local realist' correlator $P_{lr}(a,b)$ is then defined as the weighted sum:

$$P_{lr}(a,b) = \Sigma_\lambda w(\lambda) A(a,\lambda) B(b,\lambda)\,. \tag{II.4.4}$$

For the case where the frames are equal we obtain the equality $A(a,\lambda) = -B(a,\lambda)$. To get the required inequality Bell introduced an arbitrary third frame c and considered the expression:

$$\begin{aligned} &P_{lr}(a,b) - P_{lr}(a,c) \\ &= -\Sigma_\lambda w(\lambda)[A(a,\lambda)A(b,\lambda) - A(a,\lambda)A(c,\lambda)] \\ &= -\Sigma_\lambda w(\lambda)[1 - A(b,\lambda)A(c,\lambda)](A(a,\lambda)A(b,\lambda))\,, \end{aligned}$$

where we have multiplied the second term in the first line with $A(b,\lambda)^2 = 1$ and taken out an overall factor equal to the first term. This yields the second line, where we have a factor in square brackets and one in parenthesis. The factor in square brackets is always larger or equal to zero, whereas the factor in parenthesis is either plus or minus one, and may depend on λ. The sum over λ may be over terms with alternating signs. Therefore, if we plainly set all these signs to minus one, then the right-hand side is a sum of only positive terms and the result is larger or equal than the right-hand side of the equation as it stands. And that is where the inequality comes in, we obtain a bound for the absolute value of the left-hand side:

$$|P_{lr}(a,b) - P_{lr}(a,c)| \leq \Sigma_\lambda w(\lambda)[1 - A(b,\lambda)A(c,\lambda)]\,, \tag{II.4.5}$$

which yields the Bell inequality:

$$|P_{lr}(a,b) - P_{lr}(a,c)| \leq 1 + P_{lr}(b,c)\,. \tag{II.4.6}$$

We see that the inequality involves three classical correlators and three frames that can be chosen independently.

Quantum violates the bound. The fundamental issue is now whether we can arrange a set of quantum measurements that yield correlators that may violate this inequality. If we succeed, those measurement outcomes could not have been obtained from a theory with hidden variables. It is not hard to find a simple example, let us return to Figure II.4.5(d) for which we already calculated that $P(a,b) = -\cos 2\beta$. Let us choose $a = Z, b = X$,

and $c = (X + Z)/\sqrt{2}$ right in between a and b, then we obtain $P(a, b) = -\cos\frac{\pi}{2} = 0$ and $P(a, c) = P(b, c) = -\cos\frac{\pi}{4} = -0.71$. Substitution of these numerical values in equation (II.4.6) shows that the inequality is violated indeed: $0.71 \not\leq 0.29$!

In conclusion we may say that quantum theory is clear about what to expect, and the really big question was to 'just perform' such experiments. And that is what we turn to next.

Hidden no more

The history of EPR experiments performed since Bell published his inequalities is interesting on its own, because it was immensely hard to actually do the experiment in a way that would satisfy all critics. Indeed, as the stakes were so high all experiments were analyzed with the highest conceivable level of scientific scrutiny.

There were always new loopholes that the experimenters had to try and eliminate, and probably there always will remain some far-fetched loopholes for example questioning whether the experimenters have a free will to really choose the settings randomly etc. Fortunately, over the last few decades impressive progress has been made, and experiments have so much improved that it appears that Einstein-Bohr debate is finally settled and that local realism seems no longer a tenable alternative for quantum theory.

And it is for that reason that only in 2022 the achievements were given the highest degree of recognition as the Nobel prize was awarded to three pioneers who successively developed the experimental set-ups that provided full proof evidence that the hidden variable theories implementing local realism were no longer feasible. The award went to Frenchman Alain Aspect, the American John F. Clauser and the Austrian Anton Zeilinger, 'for experiments with entangled photons, establishing the violation of Bell inequalities and pioneering quantum information science.'

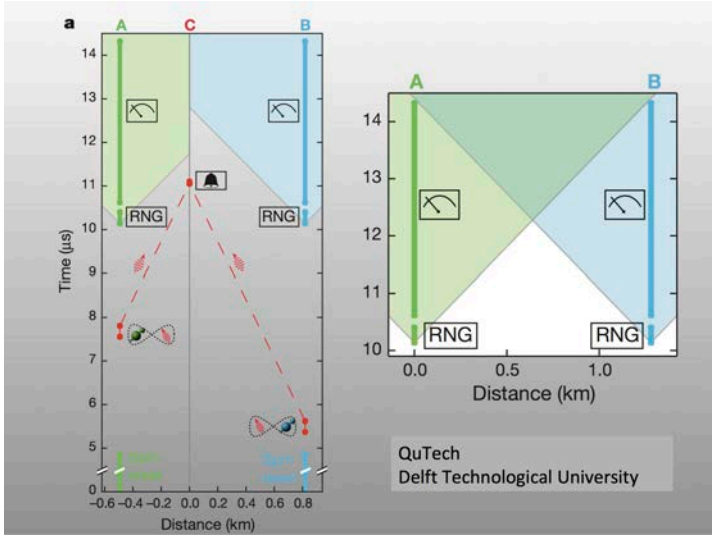


Figure II.4.6: *The Delft Experiment.* The setup of the 2015, loophole-free Bell inequality violation experiment, at Delft Technological University. The measurement stations A and B are 1.7 km apart, ensuring that the measurements are indeed spacelike separated and causally disconnected. (R. Hanson et al. *Nature*, Vol 526, 2015)

The Delft experiment. One of the more recent experiments is the 'loophole-free Bell inequality violation experiment' performed in 2015 by Ronald Hanson's group at the Delft Technological University in the Netherlands. It uses two electron spins in the maximally entangled antisymmetric state

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|1,-1\rangle - |-1,1\rangle)\,.$$

We sketched the setup of the experiment in Figure II.4.6. It involves three stations A, B, and C. In A and B two electrons are prepared in the entangled state. First each of them emits a photon so that the photon and electron are entangled. The photons are then sent through an optical fiber to station C, where they are measured in a clever way

so that that measurement can be used to verify that the electrons are indeed in the desired entangled state given above. This verification of the state to be measured is one of the loopholes that has weakened earlier attempts to corner the hidden variables option. The entangled electrons enter measurement devices in A and B, where independently a random choice between two distinct polarization directions is made for each of them. In A one chooses the observable a equal either Z or X, and in B the observable b being either $(-\mathrm{Z}+\mathrm{X})/\sqrt{2}$ or $(-\mathrm{Z}-\mathrm{X})/\sqrt{2}$. The stations A and B are 1.7 km apart, and therefore the choices and measurements are space-like separated, implying that there can't be any causal relation between them. This is indicated in the figure where the future light cones of the random choice events and measurement events at A and B are drawn, and one sees that they are outside each other's future light cones indeed. And this was another loophole that hampered earlier experiments. So this experiment really closes both the preparation and locality loopholes simultaneously and that leaves little room for the hidden variables scenario to survive. Again, one can calculate a bound on a weighted average S of the product of measurement outcomes x in A and y in B where,

$$S = |\sum_{\mathrm{ab}} \langle\psi|\mathrm{a}\otimes\mathrm{b}|\psi\rangle| \,.$$

The classical bound respecting local realism can be shown to yield $S \leq 2$, whereas the quantum value can be calculated giving $S = 2\sqrt{2} \simeq 2.83$. The highly sophisticated 2015 Delft experiment of Ronald Hanson et al. measured a total of 245 trials over a period of 18 days; this yielded an average value 2.42 with a standard deviation of 0.2.

Conclusion. We conclude that spooky action at a distance is just there and we have to live with it. Quantum weirdness is not fake; it is rock solid! It turns out to be a blessing in disguise, because it implies the spectacular possibility of quantum teleportation, to which we will turn after we have described a second experiment that also refutes the idea of local realism.

A decisive three photon experiment

There is one more experiment on entangled states that I like to describe in some more detail. It is a wonderfully conceived and designed experiment, which in a sense is so clean and therefore easy to understand, that I think it really gave a final blow to the idea of local realism and hidden variables. It is called the Greenberger–Horne–Zeilinger or GHZ experiment[2] and involves *three* (in fact even four) photons in a maximally entangled state. At first makes it may look dauntingly complicated, but the prediction is so radically unambiguous, and the reasoning so straightforward that it really is a litmus test on the matter of local realism. The answer is a clean yes or no, and does not involve a bound that has to be violated. In this experiment the outcomes predicted by the quantum hypothesis on the one hand and local realism on the other are mutually exclusive and that makes this experiment so powerful and attractive. It brings the inner workings of quantum theory to the surface. The results unambiguously prove the existence of entanglement and therefore of quantum nonlocality.

To give you an idea of the experimental setup, we have reproduced the schematic in Figure II.4.7. From a photon source maximally entangled pairs are generated, each member goes through a beamsplitter and we end up with basically four entangled photons. One of the photons is used as a trigger, and if the four detectors fire simultaneously, one knows that the three entering in the three main detectors are in a maximally entangled GHZ-state. These three photons can be analyzed in detectors *det 1, det 2* and *det 3*. The detectors are space-like separated, meaning that the measurements cannot influence each other in a causal way, and they are designed such that you can

[2] The setup of the experiment was introduced in a paper in 1989 by Greenberger, Horne, and Zeilinger and the experiments were carried out by a European collaboration of Pan, Bouwmeester, Danielli, Weinfurter and Zeilinger in 2000 (Nature, Vol 403, 2000).

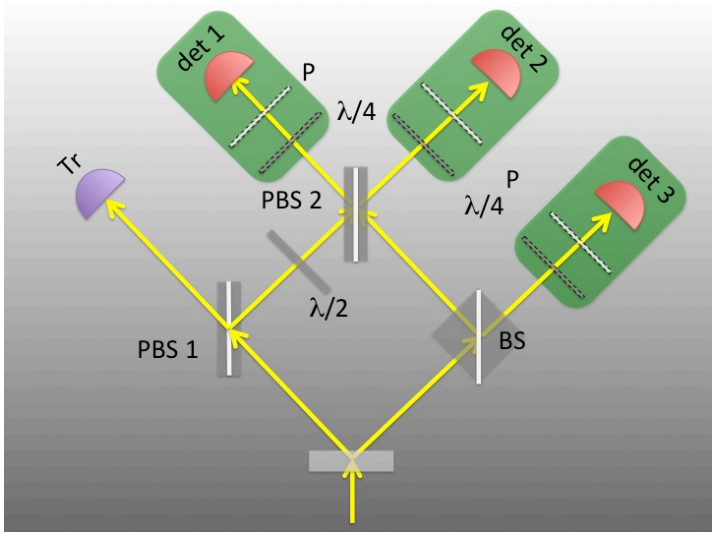


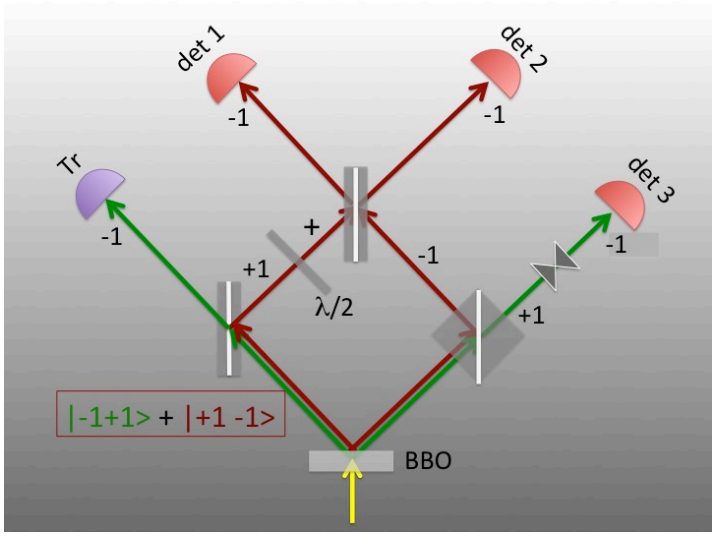
Figure II.4.7: *The GHZ experiment.* This exploits three entangled particles to unambiguously demonstrate that quantum theory violates local realism, thereby closing the door on the famous Bohr–Einstein debate. (Source: Nature, Vol 403, 2000)

switch between three different polarization bases, in particular the X-basis with eigenstates $|\pm\rangle$ and the Y-basis with eigenstates $|\mathrm{L/R}\rangle$, and the Z-basis with eigenstates $|\pm 1\rangle$, the measurement outcomes can be either $+1$ or -1. If the detectors just are in the Z-basis, you can see how the entangled state is actually prepared by the array of beam splitters and the $\lambda/2$ wave plate. The criteria for data selection is (i) that the trigger (detector) selects the events with $\lambda_z = -1$ and (ii) that indeed all four detectors pick up a simultaneous signal. These criteria can only be met in two distinct cases which we have depicted in the two figures II.4.8.[3]

Let us now analyze the quantum predictions for the experiment which starts with the three-photon GHZ state:

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|+1,+1,+1\rangle + |-1,-1,-1\rangle\,. \qquad \text{(II.4.7)}$$

[3]To be precise detector *det 3* is turned 60 degrees to invert the readout ($= 1 \leftrightarrow -1$).



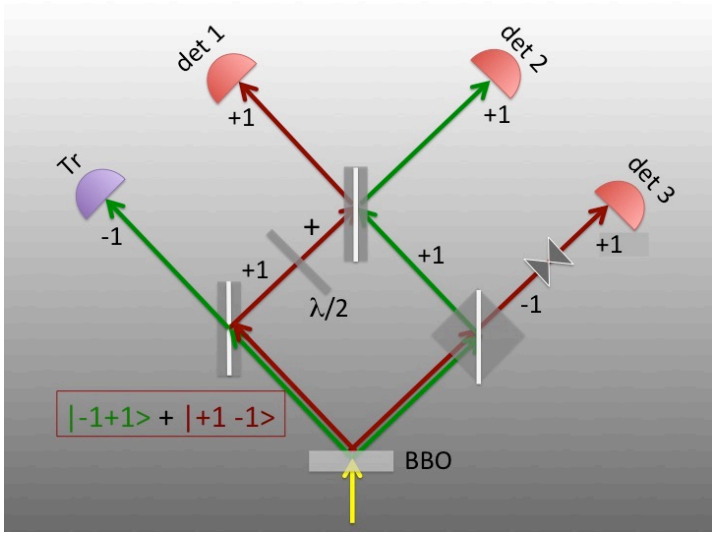


Figure II.4.8: *Contributions to GHZ.* The diagrams show the only two possible contributions to the three (or four) photon entangled state with trigger on -1, to a ZZZ measurement. (Source: Nature, Vol 403, 2000)

We can now express this state in various different bases, and GHZ proposed to study a sequence of four measurements with the detectors *det 1, det 2* and *det 3* in the following order first the cyclic variations YYX, YXY, XYY and finally an XXX measurement. Knowing the result of the

first three measurements both the quantum-adepts and the local realist followers can take that data, turn their respective cranks and come out with a unique prediction for the possible outcomes of the fourth experiment and their probabilities. The beauty of this experiment is that opposing schools of thought come out with mutually exclusive predictions! So it is a real 'yes or no' for quantum versus local realism.

So let us see how the quantum analysis goes, and it is basically what we have been doing before only a little more of it. To determine the various possible measurement outcomes and the probabilities we have to rewrite the GHZ-state in the other bases, and because we know the linear combinations this is a matter of making the appropriate substitutions in the expression (II.4.7).

$$
\begin{aligned}
|+1\rangle &= \frac{1}{\sqrt{2}}(|+\rangle + |-\rangle)\,, \\
|-1\rangle &= \frac{1}{\sqrt{2}}(|+\rangle - |-\rangle)\,, \\
|+1\rangle &= \frac{1}{\sqrt{2}}(|\mathrm{L}\rangle + |\mathrm{R}\rangle)\,, \\
|-1\rangle &= \frac{\mathrm{i}}{\sqrt{2}}(|\mathrm{L}\rangle - |\mathrm{R}\rangle)\,.
\end{aligned}
$$

So for example in the YYX experiment we would encounter the state:

$$
|\psi\rangle = \frac{1}{2}(|\mathrm{R},\mathrm{L}.+\rangle + |\mathrm{L},\mathrm{R},+\rangle + |\mathrm{L},\mathrm{L},-\rangle + |\mathrm{R},\mathrm{R},-\rangle)\,. \quad \text{(II.4.8)}
$$

let us make some observations on this state. The probability of finding a $+1$ or -1 result for any of the three photons is 50% meaning that it is maximally random: it is like throwing with a fair coin. Next note that the outcomes of each possible pair out of the three photons also has equal probabilities: so say for the first two detectors one has the that the possible outcomes $(+1,+1),(+1,-1),(-1,+1)$, and $(-1,-1)$, each occur with 25% probability. Finally it is also clear that given the outcome of two of the measurements the third is completely fixed. If the first two give LR

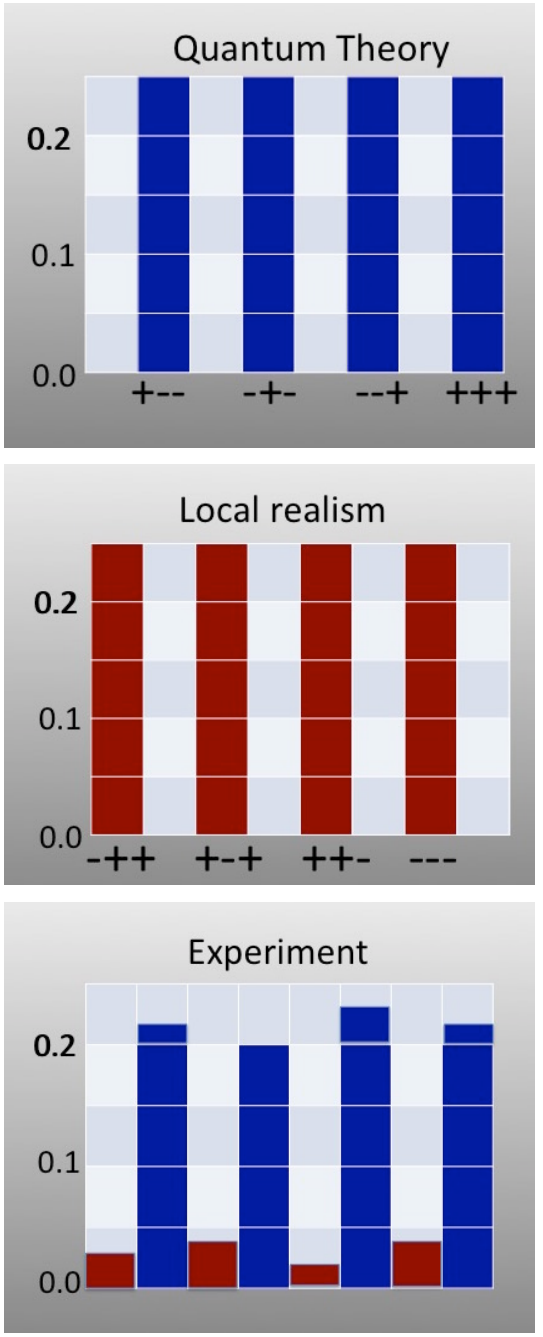


Figure II.4.9: *The decisive result.* The predictions of quantum theory (top) and local realism (middle) for the outcome of the XXX experiment are mutually exclusive. The experiment (bottom) strongly favors quantum theory. (Source: Nature, Vol 403, 2000)

or $(-1,+1)$, the third detector would have a $+$, meaning an outcome -1 for the product of the outcomes of *det 1, det 2* and *det 3*. It is clear that exactly half of the possible $2^3 = 8$ possible outcomes will occur in this experiment, and this selection is an expression of the correlations that quantum theory produces. And of course the same holds for the other three experiments in the sequence of four. Indeed for the fourth XXX experiment, we should express the state in the XXX-basis, which yields:

$$|\psi\rangle = \frac{1}{2}(|+,+,+\rangle + |+,-,-\rangle + |-,+,-\rangle + |-,-,+\rangle)\,. \tag{II.4.9}$$

The thing to note here is that the product of the measurement outcomes of the three detectors will always be $+1$, whatever the component of equation (II.4.9) is that happens to occur.

Let us now do the analysis following the local realism line of reasoning. The idea is that the setup is such that there is no causal relation between them. This means that each of the photons should carry an element of reality for both the X and Y measurements, telling us what the outcome of such a measurement would be. Let us call these elements which are just numbers, x_i and y_i where $i = 1,2,3$ labels the detector, where these can only equal ± 1. If we now look at the possible outcome of the XYY measurement and its permutations, each of the photons can carry only one particular x_i and y_i, which should fit all three possibilities in (II.4.8). This leads for the first three measurements to the three equations:

$$y_1 y_2 x_3 = -1\ ;\ y_1 x_2 y_3 = -1\ ;\ x_1 y_2 y_3 = -1\,. \tag{II.4.10}$$

The neat thing is that the solution of these three equations completely fixes the product $x_1 x_2 x_3$, which then of course is the local realism prediction for the outcome of the fourth (XXX) measurement. If we take the product of the three equations (II.4.10), we get that:

$$(y_1 y_2 x_3)(y_1 x_2 y_3)(x_1 y_2 y_3) = (x_1 x_2 x_3) y_1^2 y_2^2 y_3^2 = -1\,.$$

With the squares of the y_i being $+1$, we get the prediction $x_1 x_2 x_3 = -1$. This answer is exactly opposite to the quantum prediction following from equation (II.4.9), which as we already mentioned, gives for the product $x_1 x_2 x_3 = +1$! If we go back to the 8 possible measurement outcomes for the XXX experiment, this would lead to what is depicted in Figure II.4.9, for the predictions, and the actual measurement outcome of the experiment showing extremely strong support for quantum theory.

Quantum teleportation

Quantum teleportation provides a method for privately sending messages in a way that ensures that the receiver will know if anyone eavesdrops. This is possible because a quantum state is literally teleported, in the sense of 'beam me up Scotty': A quantum state is destroyed in one place and recreated in another. Because of the no-cloning theorem that we discussed on page 298 of Chapter II.2, it is impossible to make more than one copy of this quantum state, and as a result when the new teleported state appears, the original state must be destroyed. Furthermore, it is impossible for both the intended receiver and an eavesdropper to have the state at the same time, which helps make the communication secure.

Quantum teleportation takes advantage of the correlation between entangled states as discussed in the previous sections. Suppose Alice wants to send a secure message to Charlie at a (possibly distant) location. The process of teleportation depends on Alice and Charlie sharing different qubits of an entangled state. Alice makes a measurement of her part of the entangled state, which is coupled to the state she wants to teleport to Charlie, and sends him some classical information about the entangled state. With the classical information plus his half of the entangled state, Charlie can reconstruct the teleported state. We have indicated the process in Figure II.4.10. We fol-

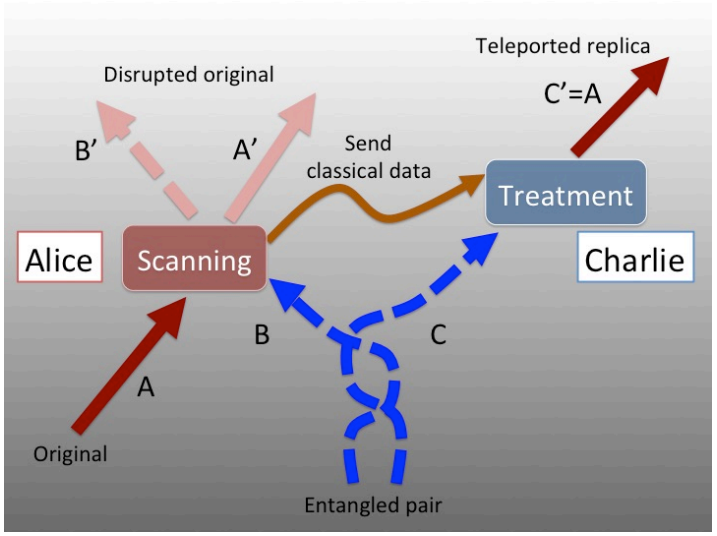


Figure II.4.10: *Quantum teleportation.* Teleportation of a quantum state using an entangled pair, as proposed by Bennett et al. in 1993. An explanation is given in the text.

low the method proposed by Bennett et al. in 1993, and first realized in an experimental setup by Zeilinger's group in 1997. In realistic cases the needed qubit states are typically implemented as left- and right-handed polarized light quanta (i.e. photons).

The simplest example of quantum teleportation can be implemented with three qubits. The (A) qubit is the unknown state to be teleported,

$$|\psi_A\rangle = \alpha|1\rangle + \beta|-1\rangle\,. \qquad \text{(II.4.11)}$$

This state is literally teleported from one place to another. If Charlie likes, once he has the teleported state he can make a quantum measurement and extract the same information about α and β that he would have been able to extract had he made the measurement on the original state.

The teleportation of this state is enabled by an auxiliary two-qubit entangled state. We label these two qubits B and C. For technical reasons it is convenient to represent this in a special basis consisting of four states, called Bell states, which are written as:

$$\begin{aligned} |\Psi^{(\pm)}_{BC}\rangle &= \sqrt{\frac{1}{2}}(|1_B\rangle|-1_C\rangle \pm |-1_B\rangle|1_C\rangle)\,, \\ |\Phi^{(\pm)}_{BC}\rangle &= \sqrt{\frac{1}{2}}(|1_B\rangle|1_C\rangle \pm |-1_B\rangle|-1_C\rangle)\,. \end{aligned} \qquad \text{(II.4.12)}$$

The process of teleportation can be outlined as follows (please refer to Figure II.4.10).

1. Someone prepares an entangled two-qubit state BC (the *Entangled pair* in the diagram).

2. Qubit B is sent to Alice and qubit C is sent to Charlie.

3. In the *Scanning* step, Alice measures in the Bell states' basis the combined wavefunction of qubits A (the *original* in the diagram) and the entangled state B, leaving behind the *Disrupted original*.

4. Alice sends two bits of classical data to Charlie telling him the outcome of her measurements (*Send classical data*).

5. Based on the classical information received from Alice, Charlie applies one of four possible operators to qubit C (*Apply treatment*), and thereby reconstructs A, getting a *teleported replica of the original*. If he likes, he can now make a measurement on A to recover the message Alice has sent him.

We now explain this process in more detail. In step (1) an entangled two-qubit state ψ_{BC} such as that of equation (II.4.12) is prepared. In step (2) qubit B is transmitted to Alice and qubit C is transmitted to Charlie. This can be done, for example, by sending two entangled photons, one to each of them. In step (3) Alice measures the joint state of qubit A and B in the Bell states' basis, getting two classical bits of information, and projecting the joint wave-

function ψ_{AB} onto one of the Bell states. The basis of Bell states has the nice property that the four possible outcomes of the measurement have equal probability. To see how this works, for convenience suppose the entangled state BC was prepared in state $|\Psi^{(-)}_{BC}\rangle$. In this case the combined wavefunction of the three-qubit state is

$$|\psi_{ABC}\rangle = |\psi_A\rangle|\Psi^{(-)}_{BC}\rangle = \frac{\alpha}{\sqrt{2}}\big(|1_A\rangle|1_B\rangle|-1_C\rangle - |1_A\rangle|-1_B\rangle|1_C\rangle\big) + \frac{\beta}{\sqrt{2}}\big(|-1_A\rangle|1_B\rangle|-1_C\rangle - |-1_A\rangle|-1_B\rangle|1_C\rangle\big). \quad \text{(II.4.13)}$$

If this is expanded in the Bell states' basis for the pair AB, it can be written in the form:

$$|\psi_{ABC}\rangle = \frac{1}{2}\Big[|\Psi^{(-)}_{AB}\rangle(-\alpha|1_C\rangle - \beta|-1_C\rangle) + |\Psi^{(+)}_{AB}\rangle(-\alpha|1_C\rangle + \beta|-1_C\rangle) \; |\Phi^{(-)}_{AB}\rangle(\beta|1_C\rangle + \alpha|-1_C\rangle) + |\Phi^{(+)}_{AB}\rangle(-\beta|1_C\rangle + \alpha|-1_C\rangle)\Big]. \quad \text{(II.4.14)}$$

We see that the qubit pair AB has equal probability to be in the four possible states $|\Psi^{(-)}_{AB}\rangle$, $|\Psi^{(+)}_{AB}\rangle$, $|\Phi^{(-)}_{AB}\rangle$ and $|\Phi^{(+)}_{AB}\rangle$.

In step (4), Alice transmits two classical bits to Charlie, telling him which of the four basis functions she observed. Charlie now makes use of the fact that in the Bell basis there are four possible states for the entangled qubit that he has, and his qubit C was entangled with Alice's qubit B before she made the measurement. In particular, let $|\phi_C\rangle$ be the state of the C qubit, which from equation II.4.14) is one of the four states:

$$|\phi_C\rangle = \begin{pmatrix}-\alpha\\-\beta\end{pmatrix}; \begin{pmatrix}-\alpha\\\beta\end{pmatrix}; \begin{pmatrix}\beta\\\alpha\end{pmatrix}; \text{ and } \begin{pmatrix}-\beta\\\alpha\end{pmatrix}.$$

In step (5), based on the information that he receives from Alice, Charlie selects one of four possible operators F_i and uses it to measure the C qubit. There is one operator F_i for each of the four possible Bell states, which are respectively:

$$F = -\begin{pmatrix}1&0\\0&1\end{pmatrix}; \begin{pmatrix}-1&0\\0&1\end{pmatrix}; \begin{pmatrix}0&1\\1&0\end{pmatrix}; \text{ and } \begin{pmatrix}0&1\\-1&0\end{pmatrix}. \quad \text{(II.4.15)}$$

Provided Charlie has the correct classical information and an intact entangled state he can reconstruct the original A qubit by measuring $|\phi_C\rangle$ with the appropriate operator F_i.

$$|\psi_A\rangle = \alpha|1\rangle + \beta|-1\rangle = F_i|\phi_C\rangle. \quad \text{(II.4.16)}$$

By simply multiplying each of the four possibilities it is easy to verify that as long as his information is correct, he will correctly reconstruct the A qubit $\alpha|1_A\rangle + \beta|-1_A\rangle$.

We stress that Charlie needs the classical measurement information from Alice. If he could do without it the teleportation process would violate causality, since information could be transferred instantaneously from Alice to Charlie. That is, when Alice measures the B qubit, naively it might seem that because the B and C qubits are entangled, this instantaneously collapses the C qubit, sending Charlie the information about Alice's measurement, no matter how far away he is. To understand why such instantaneous communication is not possible, suppose Charlie just randomly guesses the outcome and randomly selects one of the four operators F_i. Then the original state will be reconstructed as a random mixture of the four possible incoming states $|\phi_C\rangle$. This mixture does not give any information about the original state $|\psi_A\rangle$. The same reasoning also applies to a possible eavesdropper, conveniently named Eve. If she manages to intercept qubit (C) and wants 'to measure it' before Charlie does, without the two bits of classical information, she will not be able to recover the original state. Furthermore she would affect that state. If Charlie somehow gets the mutilated state, he will not be able to reconstruct the original state A. Security can be achieved if Alice first sends a sequence of known states which can be checked by Charlie after reconstruction.

Superposition The strange thing about a qubit in comparison with its digital precursor is the fact that it can be in a state that is a 'superposition' of the '1' and the '0' state. This is possible because of the all-important *linear superposition principle* which is a basic ingredient of quantum theory. As a consequence of quantum information processing, the manipulation of qubits, i.e. changing their states by having them interact, is like doing parallel processing on a large scale. The exceptional power of the quantum computers of the future is a reflection of the ability to directly work with these linear superpositions. Here is an analogy that may help you understand why this is so. Imagine you would like to make a street map of a city to find the shortest route from point P on one side of town to point Q on the opposite side. As a single being you go and walk in the right direction, and to find the shortest route you should walk in principle all the possible routes that bring you from P to Q and compare their lengths. Parallel processing would mean that you hire a bunch of students to independently and simultaneously take different paths from P to Q. This certainly will save time. But now imagine that some *Dr Ghetto Blaster* comes along with a device which produces lots of sound at point P and his business partner *Dr Ghetto Digest* sits down at point Q with an impressive highly sophisticated listening device. He turns the machine on and in no time has reconstructed the street map. Imagine! The remarkable thing is that this is in principle possible because sound as an agent always takes all possible paths through town simultaneously, and interferes with itself on every corner, and all that information is encoded in the changes of the signal that we would receive in Q. It probes the street plan not sequentially but in parallel. A fashionable version of this story is to say that you can hear the shape of a remote drum if somebody is playing it, or that you can hear the shape of a tin roof by listening to the rain pouring on it. This is so because there are many ticks and every tick in a sense 'contains' all frequencies and therefore these examples are classical analogues and show the potential power of the linear superposition principle. □

If the original and reconstructed sequence are perfectly correlated, then that guarantees that Eve is not interfering. Note that the no-cloning theorem is satisfied, since when Alice makes her measurement she alters the state ψ_{A} as well as her qubit B. Once she has done that, the only hope to reconstruct the original ψ_{A} is for her to send her measurement to Charlie, who can apply the appropriate operator to his entangled qubit C.

The quantum security mechanism of teleportation is based on strongly correlated, highly non-local entangled states. While a strength, the non-locality of the correlations is also a weakness. Quantum correlations are extremely fragile and can be corrupted by random interactions with the environment, i.e. by decoherence. As we discussed before, this is a process in which the quantum correlations are destroyed and information gets lost. The problem of decoherence is the main stumbling block in making progress towards large-scale development and application of quantum technologies. Nevertheless, the research group of Gisin et al. at the University of Geneva demonstrated teleportation over a distance of 550 meters using the optical fiber network of *Swisscom* in 2006.

A n important next step would be the construction of a network of quantum devices with links along which entangled states can be created and quantum information teleported securely. In 2022 the first successful steps were reported by the *QuTech* group of Hanson in Delft.

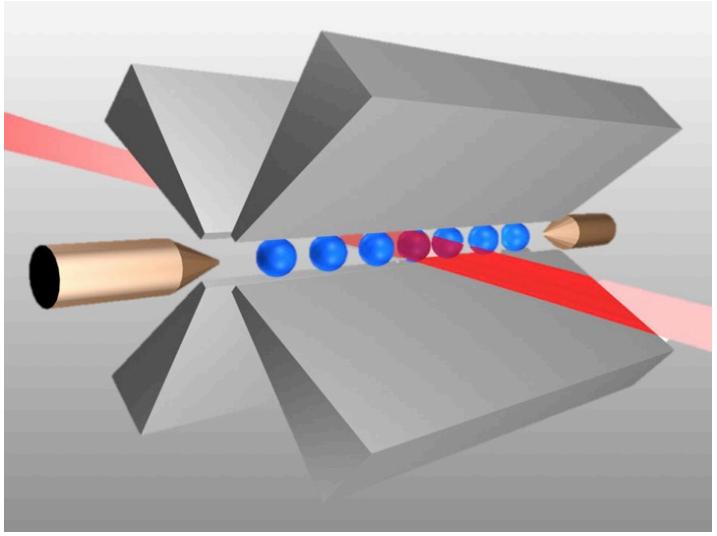


Figure II.4.11: *Trapped ions.* Ions trapped in a linear optical lattice. *(IQO Insbruck)*

Quantum computation

Quantum computation is performed by setting up controlled interactions that cause non-trivial dynamics and successively couple individual qubits together and generate a time evolution of the quantum state in a predetermined manner. And moreover ensuring that no other interactions take place that could corrupt the computation. A multi-qubit system is first prepared in a known initial state, representing the input to the program. Then interactions are switched on by applying forces, such as magnetic fields, that determine the direction in which the wavefunction rotates in its state space. Thus a quantum program is just a sequence of unitary operations that are externally applied to the initial state. This is achieved in practice by a corresponding sequence of quantum gates. When the computation is done measurements are made to read out the final state. Measurements are non-unitary operations that can also be part of the process.

Quantum computation is essentially a form of analog computation. A physical system is used to simulate a mathematical problem, taking advantage of the fact that they both obey the same equations. The mathematical problem is mapped onto the physical system by finding an appropriate arrangement of magnets or other fields that will generate the proper equation of motion. One then prepares the initial state, lets the system evolve, and reads out the answer. Analog computers are nothing new. For example, Leibnitz built a mechanical calculator for performing multiplication in 1694, and in the middle of the twentieth century, because of their vastly superior speed in comparison with digital computers, electronic analog computers were often used to solve differential equations.

Then why is quantum computation special? The key to its exceptional power is the massive parallelism at intermediate stages of the computation. Any operation on a given state works simultaneously on all basis vectors and thus also on entangled states. The physical process that defines the quantum computation for an n qubit system thus acts in parallel on a set of 2^n complex numbers, and the phases of these numbers (which would not exist in a classical computation) are important for determining the time evolution of the state. When the measurement is made to read out the answer at the end of the computation we are left with the n-bit output and the phase information is lost.

Because quantum measurements are generically probabilistic, it is possible for the 'same' computation to yield different 'answers', e.g. because the measurement process projects the system onto different eigenstates. This can require the need for error correction mechanisms, though for some problems, such as factoring large numbers, it is possible to test for correctness by simply checking the answer to be sure it works. It is also possible for quantum computers to make mistakes due to decoherence, i.e. because of essentially random interactions between the quantum state used to perform the computation and the environment. This also necessitates error correction mechanisms.

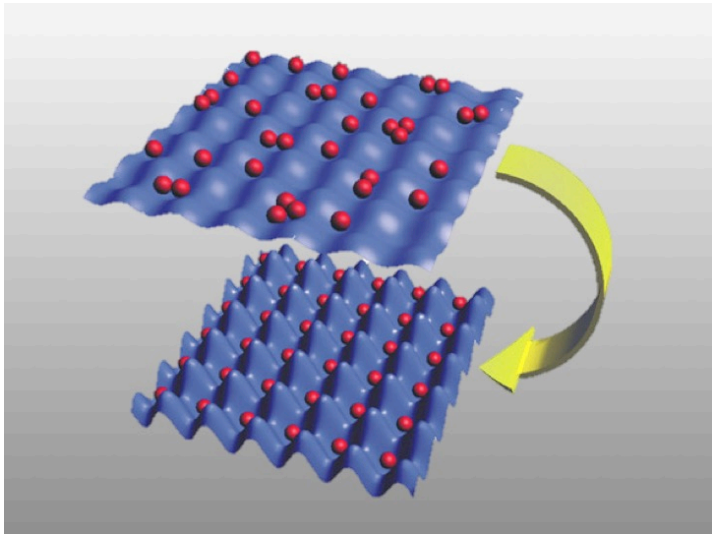


Figure II.4.12: *Optical lattice.* Atoms can be manipulated in a linear optical lattice. *(IQO Innsbruck)*

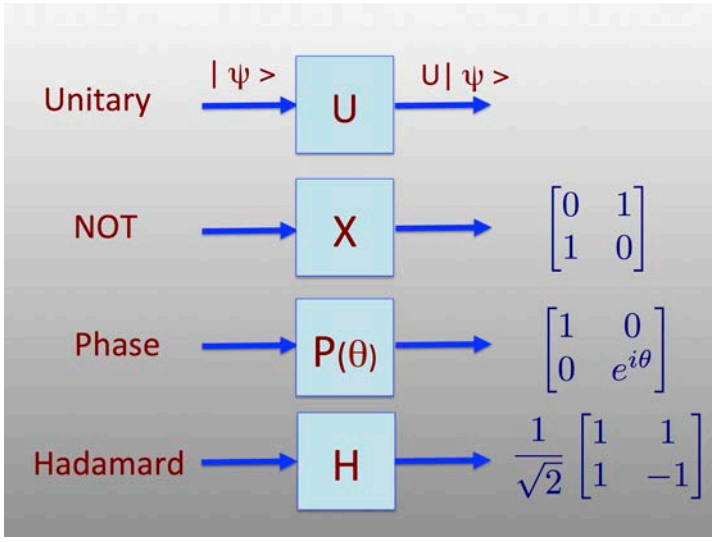


Figure II.4.13: *Gates.* Some standard one-bit quantum gates.

The problems caused by decoherence are perhaps *the* central difficulty in creating realistic physical implementations of quantum computation. These can potentially be overcome by constructing quantum systems where states are not encoded locally, but rather globally, in terms of topological properties of the system that cannot be disrupted by external (local) noise. This is called *topological quantum computing*. This interesting possibility arises in certain two-dimensional physical media which exhibit *topological order*, referring to states of matter in which the essential quantum degrees of freedom and their interactions are topological (see also Chapter III.3).

Quantum gates and circuits

In the same way that classical gates are the building blocks of classical computers, quantum gates are the basic building blocks of quantum computers. A gate used for a classical computation implements binary operations on binary inputs, changing zeros into ones and vice versa. For example, the only non-trivial single bit logic operation is NOT, which takes 0 to 1 and 1 to 0. In a quantum computation the situation is quite different, because the states of qubits live in a two-dimensional Hilbert space and they represent complex superpositions of 0 and 1. This was discussed in considerable detail in Chapter II.1.

Single qubit gates. The set of allowable single qubit operations consists of unitary transformations corresponding to 2×2 complex matrices U such that $\mathrm{U}^\dagger\mathrm{U} = 1$. The corresponding action on a single qubit is represented in a circuit as illustrated in Figure II.4.13.

Some quantum gates have classical analogues, but many do not. As we mentioned, the X operator is the quantum equivalent of the classical NOT gate, and serves the function of interchanging spin up and spin down. In contrast, the Z operator rotates the relative phase of the two-component wavefunction by 180 degrees and has no classical equivalent.

Let us briefly discuss the typical one-qubit logical gates of

Figure II.4.13. First the NOT gate,

$$X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

as we mentioned this is the quantum equivalent of the classical NOT gate and acts by interchanging $|1\rangle$ and $|-1\rangle$. The next one is

$$P(\theta) = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\theta} \end{pmatrix}.$$

The $P(\theta)$ operation is called the phase gate, since it changes the relative phase by θ degrees.

The third gate is the so-called Hadamard gate H,

$$H = \sqrt{\frac{1}{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

which creates a superposition of the basis states: $|\pm 1\rangle \Rightarrow |\pm\rangle$. In other words it flips between the Z- and the X-frames.

The general purpose of a quantum computer is to transform an arbitrary n-qubit input into an n-qubit output corresponding to the result of the computation. In principle implementing such a computation might be extremely complicated, and might require constructing quantum gates of arbitrary order and complexity.

Universal gate sets. Fortunately, it has been shown that the transformations needed to implement a universal quantum computer can be generated by a simple – so-called universal – set of elementary quantum gates, for example involving a well-chosen set of one- and two-qubit gates. Single qubit gates are unitary matrices with three real degrees of freedom. If we allow ourselves to work with finite precision, the set of all gates can be arbitrary well approximated by a small, well-chosen set. There are many possibilities – the optimal choice depends on the physical implementation of the qubits.

From the perspective of experimental implementation, a convenient two-qubit gate to use is the CNOT gate we have discussed before, see Figure II.1.17. The combination of the CNOT, the $P(\pi/4)$ and the Hadamard gate forms for example a universal set.

Shor's algorithm

Prime factoring. An algorithm is not an equation; it is more like an operational set of steps – a procedure – that is *guaranteed* to lead to a desired result. So it usually does involve equations and a mathematical proof. For the Shor algorithm the problem is to factor a large number, say of about 800 or 1000 digits, into its prime factors, in most cases there are just two of them. So we have a number N that can being written in a unique way as a product of two prime numbers a and b. One way to do this is just by trial and error. In fact by checking one after the other whether $2, 3, 5, \ldots$ is a divisor of the number N. And this you may do by a simple subtraction scheme à la Euler, where you keep subtracting the candidate divisor and look whether you indeed hit zero. As we have argued in Chapter I.3, such schemes end up being extravagantly costly in the time it takes to actually factor a really big number. That time is significantly longer than the age of the universe and that should not surprise you. The one thing it makes at least clear is that patience will not suffice. The time dependence on N if one uses conventional digital computers is typically exponential. The showcase example of why quantum computers are indeed fundamentally different, and for a task like this one far superior, is the *Shor factorization algorithm* which is a quantessential algorithm, because it exploits non-commutativity of operators in a clever way.

The MIT applied mathematics professor Peter Shor proposed the algorithm in 1994 and was co-recipient of the 2023 Breakthrough Prize in Fundamental Physics.

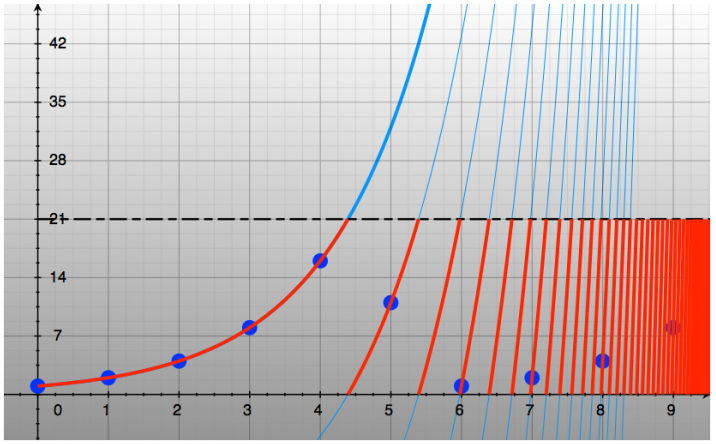


Figure II.4.14: *The periodic function.* We have displayed $f(x) = c^x \bmod N$ (red curve). The blue points represent the (discrete) periodic function over the integers. We have chosen $c = 2$ and $N = 21$. The period equals 6.

The algorithm. The algorithm for factorization consists of three steps.
(i) construct a particular periodic function modulo N,
(ii) determine the period of that function,
(iii) given (i) and (ii) one can use an Euler method for finding largest common divisors to find the factors a and b such that $ab = N$.

(i) Construction of the periodic function. Choose an integer c and consider the function[4]

$$f(m) = c^m \bmod N\,,\ m \text{ integer}\,. \tag{II.4.17}$$

(ii) One can show that this function is periodic with a period we call r, so,

$$f(m + r) = f(m)\,. \tag{II.4.18}$$

After substitution of f on both sides, it then follows that

$$c^r = 1 \bmod N \ \rightarrow\ c^r - 1 = sN\,,$$

where s is some integer. Now rewrite the left-hand side as:

$$(c^{r/2} + 1)(c^{r/2} - 1) = sN\,,$$

where we need r to be even for the factors to be integers. If r happens to be odd, one has to restart by choosing a different value for c and start all over.
(iii) The next step is to find the greatest common divisor of the individual factors on the left with N, after which one obtains the prime factors a and b of N. This last step can be done with an Euler subtraction scheme.

The hard part of this solution method is to find the period r of the function $f(m)$ because this r may be of order N itself. Determination can be done using a fast or integer Fourier transform of $f(m)$.

As we discussed wavefunctions, and non-commuting operators as hallmarks of quantum theory it is maybe nice to paraphrase this hard side of the problem and to see that quantum measurement is the clue. Firstly think of the function $f(m)$ as a wavefunction on a one-dimensional lattice corresponding to the natural numbers $0, 1, 2, 3, \ldots$. Now we also have discussed a momentum operator P which translates the position variable by one unit. And acting on the function it acts like $P\ f(m) = f(m + 1)$. Because of the periodicity of $f(m)$ we also have the relation $P^r f(m) = f(m + r) = f(m)$ from which we conclude that $P^r = 1$. From which it follows that the eigenvalues of the P operator are $p = e^{2\pi i s/r}$ with $s = 0, 1, \ldots, r$. In other words doing a measurement of the momentum of the state described by the wavefunction $f(m)$ tells us basically what r is![5] What we end up with is a periodic function with a support of r points on a circle and dual to that the momentum sample space also consisting of r points. This is of course due to the periodicity of the function. I recall the statement about the relation between the sample spaces of position versus momentum operators. A line is dual to a line. If the

[4] The number $m = M \bmod N$ is obtained by subtracting N from M until a number between 0 and N is obtained, which is the number m. In other words $M = m + kN$ for some k with $0 \leq m < N$.

[5] One may need more than one measurement, but one can check that rather easily.

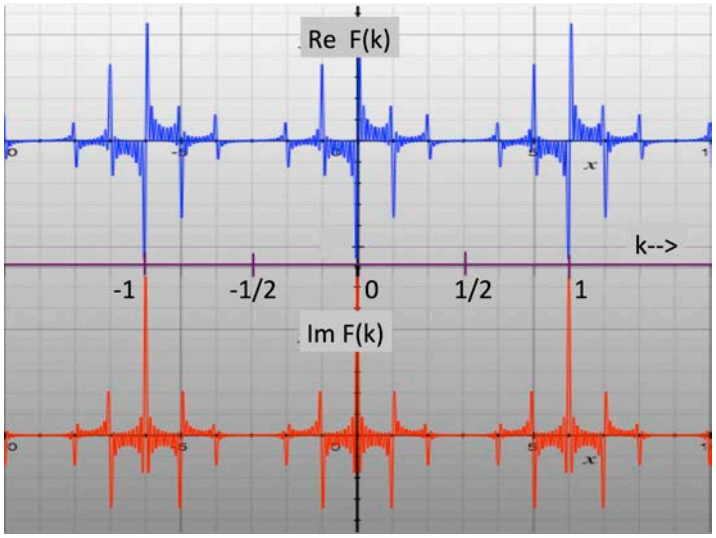


Figure II.4.15: *The Fourier transformed function* $F(k)$. We have displayed $Re\ F(k)$ (blue curve) and $Im\ F(k)$ (red curve). The peaks at multiples of $1/6$ stand out clearly, even in this crude 'iPhone' approximation causing some noise. This means that the periodicity of the original function $f(m)$ (the blue dots in the previous figure) would be 6.

x-space is infinite discrete then the sample space of the dual momentum is a angle or a circle, by bringing in the periodicity only a set of r points on the circle is left corresponding to the corners of a polygon. And in that case the P and X sample spaces are again the same. There are basically two identical polygons and there is a unitary transformation between the frames that correspond to the sets of eigenvectors corresponding to the eigenvalues. Stated differently the problem of factoring is to a large extent finding the right polygon hidden in the circle and indeed there are many (a countable infinity) to choose from.

The fast Fourier transform of a function $F(n)$ is defined as:

$$F(k) = \sum_m f(m) e^{2\pi i k m}\,, \qquad \text{(II.4.19)}$$

which combined with the fact that $f(n)$ has a period r leads to a powerful conclusion on the function $F(k)$. For the function (II.4.17) it leads to the strong condition:

$$e^{2\pi i k r} = 1 \;\rightarrow\; k = s/r\,;\; s = 1, \ldots, r\,.$$

What this means is that we ask for the transformation of a (wave) function $f(x)$ on a one-dimensional infinite lattice, from the position state basis to a momentum state basis. We know that the momentum values for an infinite discrete space correspond to an angle $0 \leq \theta \leq 2\pi$ where in our case $\theta = 2\pi\, k = 2\pi\, s/r$. So what we learn is that the function f involves only r different momentum states. The fast Fourier transform just measures the momentum and determines the component of that momentum eigenstate. The magnitude of that component is not so relevant as what the actual allowed momenta are. So the momentum state is almost everywhere zero except in points that correspond to the corners of a polygon with r sides where they have the value $F(k)$.
Wouldn't it be fun to find an example where we would be left with a pentagon, what do I say, THE pentagon, in momentum space? Maybe that explains the Pentagon's interest in quantum computing and maybe they knew all along that the pentagon would play an important role somewhere....

So the data we need from the fast Fourier transform just corresponds to one or more measurements of the momentum in the state f. That will give us a value(s) $p = 2\pi s/r$ from which r can be determined. So it is now clear that quantum measurements implement an extremely efficient algorithm for fast Fourier transform on integer-valued functions. You just have to measure the non-commuting observable dual to the variable of the function, and that is the momentum. And that is the quantessence of super fast factorization.

Let us work out a simple example, and let us try to factor the number $N = 21$ with the algorithm. We first construct the function $f(x) = 2^x \bmod N$, it takes the values given in Table II.4.1. We see that the function has a period $r = 6$,

Table II.4.1: Tabulation of the function $f(x) = 2^x \bmod 21$.

x	0	1	2	3	4	5	6	7	8	9	10
f(x)	1	2	4	8	16	11	1	2	4	8	16

so we obtain the equation:

$$(2^3 + 1)(2^3 - 1) = 9 \times 7 = 21 \times s\,.$$

Now determine the largest common divisor from the factors on the left with 21:
$21 \bmod 7 = 0 \rightarrow 7$ is a factor of 21, and $21 \bmod 9 = 3 \rightarrow 3$ is a factor of 21. Thus we established the magical result that $21 = 3 \times 7$, One could say that we at least succeeded in cracking a nut by using a magnificent sledgehammer.

But to factor a 1000 digit number into two primes you will need this sledgehammer in the form of a sizable quantum computer to find the period, which after all might well be of the order of N itself!

Applications and perspectives

Quantum computation and security are challenging examples of the surprising interplay between the basic concepts of physics and information theory. If physicists and engineers succeed in mastering quantum technologies to allow for reliable and scalable qubits, it will mark an important turning point in information science with profound societal consequences. We had better get ready for an era of quantum supremacy!

Hardware developments. As we mentioned already, at present there is a lot of work in progress trying to implement quantum computing in a wide variety of ways. I will refrain from going into any detail here firstly because that calls for many different types of expertise, and furthermore the developments go so fast and still make so many unexpected turns that I would run the risk that this book would already be out-of-date before it was published. It is absolutely clear however that basically all big tech companies are actively pursuing the quantum opportunities that suits them. In principle all that is needed to make a qubit is a simple two-level quantum system that can easily be manipulated and scaled up to a large number of qubits. The first requirement is not so restrictive, and many different physical implementations of systems with a single or a few qubits have been realized, including NMR, spin lattices, linear optics with single photons, quantum dots, Josephson junction networks, ion traps and atoms and polar molecules in optical lattices.

The much harder problem that has so far limited progress toward practical computation is to couple the individual qubits in a controllable way and to achieve a sufficiently low level of decoherence. Even small local perturbations due to the environment could destroy the delicate phase information in the linear superposition of states. With respect to this problem, a promising venue has surfaced with the advent of *Topological Quantum Computing* where quantum information is stored in topological degrees of freedom that are insensitive to local perturbations and interactions, making error correction procedures simpler to implement. This way of computing involves new states of matter, that exhibit what is called topological order. In Chapter III.3 we'll say more about this. On the software side impressive progress has been made, building on the fundamental quantum algorithms we have mentioned. There is of course also the possibility of developing hybrid classical/quantum devices. Nevertheless, with the great efforts now taking place, future developments could be surprisingly fast.

The challenge of quantum software. We are in a situation that looks like the early seventies where many institutions in what still was Silicon Valley to be, started focussing on developing software for digital devices like PC's and laptops, that weren't really there yet. This major effort

was to a large extent based on the strongly held belief that a digital era was on its way where every individual would own powerful devices, to play and work with. High level languages had to be developed to allow everybody to optimally process data, whether it concerned text, pictures, symbolic manipulation or music. It turned into an unprecedented show-case of public and private research and development efforts, which resulted in the present information era which in many ways has profoundly changed the human condition.

We are now in a comparable situation with respect to quantum computing. And again, even though the hardware is still quite remote, a strong case for quantum software should be made. If we were to have quantum computers at our disposal, the question of what miracles could they possibly perform strongly depends on the software that is available. We said in the introduction to this section that there are many problems where the intrinsic massive parallelism of quantum evolution might yield dramatic speedups in computation. The point is not that a classical computer would not be able to do the same computation – after all, one can always simulate a quantum computer on a classical one – but rather the time that is needed could drastically be reduced.

As we just discussed in some detail, a most spectacular speedup is achieved by the Shor algorithm (1994) for factoring large numbers into their prime factors. Because many security keys are based on the inability for digital computers to do this, the reduction from an exponentially hard to a polynomially hard problem has many practical applications for breaking security codes and current cryptography. This means that even today, one has already to worry about how one should save sensitive information, to make sure that it cannot be easily retrieved in the near quantum future. Quantum algorithms also allow one to provide new more secure crypto-codes that in principle allow users to run programs on untrusted systems and still keeping their data secret.

Another important application is the quadratic speedup by Grover's search algorithm (1996) over conventional search algorithms, addressing for example problems like the 'traveling salesman', in which large spaces of possibilities need to be searched and compared.

Machine learning is another hot topic where the discovery of an exponential speedup for solving certain systems of linear equations has led to flurry of new developments like algorithms for core problems like data fitting and supporting vector machines.

Finally, a vital application is the efficient simulation of quantum physical and chemical systems, which at present is an extremely costly business taking up much of our supercomputer capacity. This development is of importance to fields like chemistry, material science and high-energy physics. In this area a quantum computer naturally would offer an exponential speedup, which in turn would directly feed back into the successful development of new quantum technologies. Science is time and again an incredible innovation engine, we are standing at the dawn of a new era and wonder where quantum technologies will lead us.

Further reading:

On the interpretation of quantum theory:

- *Quantum: Einstein, Bohr and the Great Debate About the Nature of Reality*
  Manjit Kumar
  Icon Press (2009)
- *The Interpretation of Quantum Mechanics*
  Roland Omnes
  Princeton University Press (1994)

On quantum computing:

- *Quantum Computing for the Quantum Curious*
  Jessica Turner et al.
  Springer Link (2021)
- *Quantum Information Theory*
  Mark M. Wilde
  Cambridge University Press (2013)
- *Quantum Computation and Quantum Information*
  Isaac Chuang and Michael Nielsen
  Cambridge University Press (2011)

Chapter II.5

Particles, fields and statistics

> In fact the smallest units of matter are not physical objects in the ordinary sense; they are forms, ideas which can be expressed unambiguously only in mathematical language.
>
> *Werner Heisenberg*

In Chapters II.1 and II.2, we mainly focussed on the qubit, because in its simplicity it was most suitable to demonstrate the quantessentials. In this chapter we turn to particles and fields. We start by discussing the one-particle Schrödinger and Heisenberg equations in more detail. Next, we turn to fields and their quantization, and explain how the resulting Hilbert space describes multiparticle states. We close the chapter with a discussion of the topological origins of indistinguishability, Pauli's exclusion. principle and the spin-statistics connection.

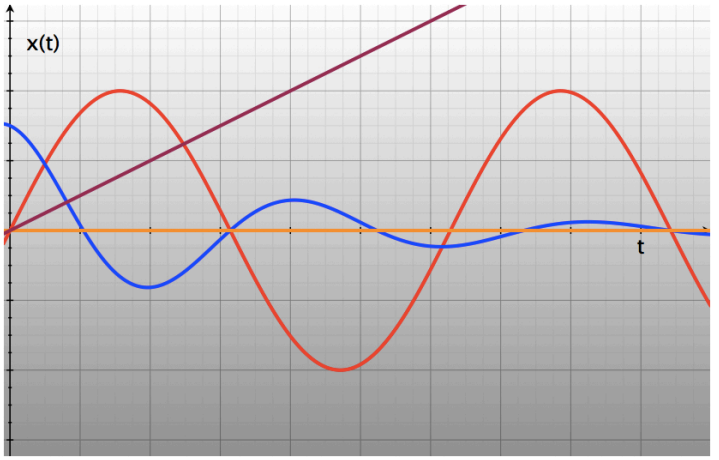


Figure II.5.1: *Moving particles.* Various particle motions as a function of time (t) in *configuration* (x) *space*. A particle successively: at rest (orange), moving with constant momentum (purple), in an oscillatory motion (red), and in a damped oscillation (blue).

Particle states and wavefunctions

Whereas the state of a single particle in classical physics is fixed by specifying its position and its velocity, i.e. by giving 6 numbers, the state of a quantum particle is specified by giving its wavefunction, a continuous function that extends over all of space. How different can the quantum world be?

Phase space. Let us consider a single particle with a given mass m and assume that it has no internal structure. In classical mechanics we specify its state by just saying what its position x and its velocity v or momentum $\mathrm{p} = \mathrm{mv}$ are. Once we fix its position and momentum at a given instant in time, Newton's laws would do the rest, given the force they completely determine the future states of the particle. The motion of the particle can be thought of as an orbit or trajectory parametrized by time in ordinary three-dimensional position or *configuration space* of the particle.

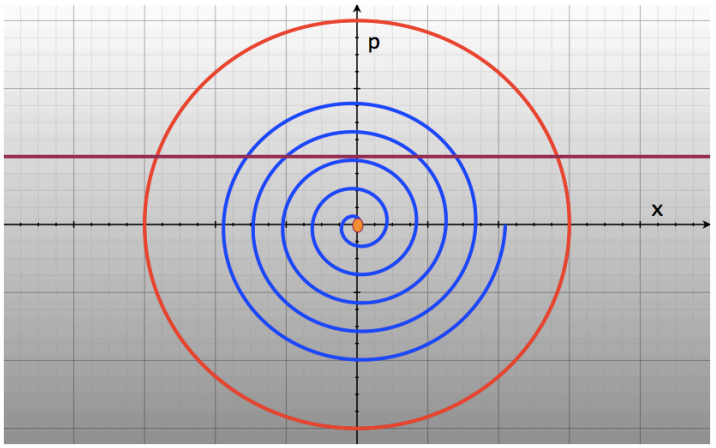


Figure II.5.2: *Particle motions.* The same particle motions as in Figure II.5.1 as a curve parametrized by t in *phase space* $(x(t), p(t))$.

In the one-dimensional case we would plot the position $x = x(t)$ with the value x on the vertical axis as a function of t along the horizontal axis. Alternatively we may think of the motion as a time parametrized curve through the combined momentum and position space which is also called the *phase space* of the particle. The phase space has twice the number of dimensions, because to the d components of the position vector one has to add the d components of the momentum vector.

We have given some examples of one-dimensional particle motions in the Figures II.5.1 and II.5.2, showing what they look like in configuration as well as phase space. So far the classical story of a particle.

Wavefunctions. The story in quantum mechanics is very different. There the state of a particle at a given time t is described by its *wavefunction* $\psi(x, t)$ which is a function that even for a single particle is defined over all of position (configuration) space.[1] Note, however, that we do not specify its velocity. If we just give the wavefunction over all of space at some initial time, then the Schrödinger equation would generate the future states given the expression for the kinetic energy and potential energy. The Schrödinger equation determines the time evolution of the wavefunction which in turn describes the particle state, and in that sense does for a quantum particle what Newton's equations did for the classical particle. We encountered this equation before in Chapter I.4 on page 158 but we will recall some of the results here for convenience. Our intuition about particles is deeply rooted in the Newtonian paradigm in that we think of a particle having a definite position a definite velocity, and that image is of course a long way from specifying some smooth function over all of space. Indeed this is nothing less then a conceptual leap that took the brightest minds a long time, first to bridge, and later to really swallow.

[1] For readers who are not already familiar with the notion of functions and what you can do with them I recommend looking at the *Math Excursion* on page 607 of Part III.

Particle-wave duality

In classical physics the particle and wave concepts are distinct and mutually exclusive. In quantum theory a particle may manifest itself in both guises. Here the concept of complementarity rears its head. The concept of a quantum particle transcends the classical distinction and appears to be both. Niels Bohr applied the wave picture to atomic orbits and obtained a discrete set of energy levels of which the lowest one is stable. A new door for fundamental physics opened up.

The vastly different framework of quantum mechanics we just outlined expresses the quantessential feature known as *particle-wave duality*. The wavefunction expresses the wave nature of a particle and the Schrödinger equation is a wave-type equation for the matter-wave that represents the particle in quantum theory. In the early days one referred therefore to quantum mechanics as 'wave mechanics.' That term sounded in the classical context rather like a

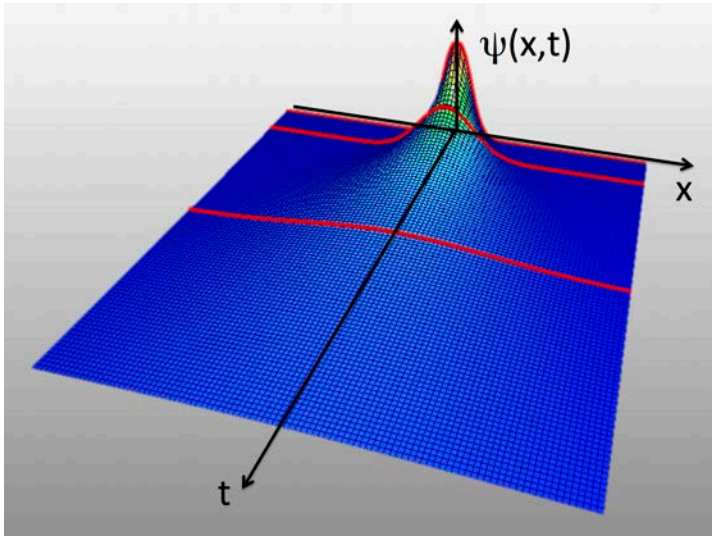


Figure II.5.3: *Particle probability density.* A quantum probability density of a particle $\Psi(x,t)$ as a function of x and t. It describes a particle at rest, well localized around the origin for $t = 0$ and then spreading out (disperse) over space as time progresses.

contradiction in terms, because in classical physics, particles and waves are fundamentally different concepts. Particles are supposed to be very much localized, while for waves the opposite holds, they typically are spatially extended. Particles can collide locally and exchange momentum and energy like billiard balls, while waves 'interact' typically by interference where the combined waves show a particular pattern of maxima and minima like water waves in a pond. We may ask what the special properties of a wave representing a particle are, or for that matter what the particle properties are of a wave, for example an electromagnetic wave.

Photons. To start with the latter, it was one of Einstein's seminal contributions to quantum theory to postulate the so-called *photon* as the quantum particle of light. Its defining properties are that this particle moves with the velocity of light, has zero mass, and an energy $E = h\nu$, where ν is the frequency of the light wave.[2] Thus a steady electromagnetic wave of a single frequency would correspond to a constant flux of particles with a fixed energy or momentum. The quantization of energy of radiation of a given frequency implied that the minimal amount of energy of a wave with frequency ν had to just be $h\nu$, and this quantization of energy was exactly what the radical postulate of Max Planck amounted to, the postulate which started off the whole quantum revolution. It was this assumption which rescued the classical black body radiation law of Rayleigh-Jeans from its demise in the high frequency domain as we pointed out in Chapter I.2.

Matter waves. It was the French physicist De Broglie who turned the relation around. He postulated the existence of matter waves: for any particle type with a mass m, the wavelength had to satisfy the relation $\lambda = h/p$, linking the wavelength to the momentum. This relation is consistent with Einstein's formula $E = h\nu$ once you realize that for a massless particle according to special relativity $E = cp$ as we pointed out in Chapter I.2, and that for a lightwave we have that $\lambda = c/\nu$.

The Bohr atom. Furthermore this picture of a matter wave was at the heart of the atomic model of Bohr, where a definite energy state of an electron would have a single wavelength but to make it periodic, it had to fit exactly on the classical circular orbit with that energy. Imposing this relation lead to the quantization of the wavelength, and thus of the momentum and therefore also to the quantization of the allowed energy for the atomic states. Bohr's picture of the atom predicted the discrete spectrum of energy

[2] This may at first sight seem problematic perhaps, because if a particle has mass equal to zero would then Einstein's own dictum – $E = mc^2$ – not decree that its energy would be zero as well? Not really, because we have to make the distinction between the *rest mass* m_0 of a particle and its *relativistic mass* m. These are related by $m^2 = m_0^2 + (p/c)^2$, showing that (i) if the momentum $p = 0$ indeed $m = m_0$, and (ii) that if $m_0 = 0$ then $m = |p|/c$. This tells us that in the latter case where the rest mass is zero, the relativistic mass is proportional to the momentum of the particle. Therefore, in relativity massless particles make complete sense and the photon is the omnipresent manifestation of that.

levels but most importantly also the existence of a lowest energy or *ground state* for the atom. The ground state corresponds to the largest wavelength that would fit on the orbit, i.e. being equal to that orbit. This point is all-important, exactly because the classical realization of an atom lacks a true ground state, the system would be unstable and the electron would fall into the nucleus in a short time, loosing energy by radiating. So the extremely stable atom as we know it in nature severely violated the laws of classical physics, and that was one of the reasons we had to give up, not just on the naive model of the atom but on the whole of classical physics! It was quantum theory that provided a fundamental understanding of the stability of matter.

Where is the particle? If a particle is represented by a wavefunction, the first question that comes mind is: 'but what about the position of the particle?' I have told you what the momentum of the particle is but where is it? Indeed, where is the particle if it is a kind of standing wave spread out around the nucleus? A perfect monochromatic wave has in principle an infinite extent. It is a periodic function like a sine or a cosine, but how can such a function ever single out any particular position for the particle? Well, you are right, it cannot.

The resolution of this tantalizing paradox has to do with the interpretation of the wavefunction and what it means to make a position measurement of a particle. We have touched on these matters already in Chapter II.2 where we learned that this comes about because of the incompatibility of different observable quantities and the frameworks that limit the degree to which questions may or may not have meaningful answers. For the moment we accept the euphemism that Niels Bohr invented for this inconvenient truth of particles being waves and *vice versa*: he called it *complementarity*. We return to these questions explicitly shortly.

The space of particle states

We extend the symbolic mathematical representation from qubit to particle states. It is profitable to also think of wavefunctions as state vectors. The square of the wavefunction defines a probability distribution of where to find the particle.

In previous chapters we looked at the space of quantum states of a system that classically corresponds to a system with a finite number of states, like an array of qubits. Now, we want to extend this discussion to a system of a particle with mass m that moves in Euclidean space. The essential difference is that the classical configuration space is now continuous.

Hilbert space heuristics. Essentially, making the step involves going from a discrete to to a continuous space and that is from a mathematical point of view a subtle matter. For that reason we will restrict ourselves here to rather heuristic arguments. If a particle could sit only in a discrete set of positions $\mathrm{x_i}(\mathrm{i}=1,\ldots,\mathrm{N})$, then of course the analysis is reduced to the one we had in the previous chapters and we would introduce a set of corresponding basis vectors $|\mathrm{x_i}\rangle$, which would be eigenvectors of the position operator X and hence satisfy the eigenvalue equation:

$$\mathrm{X}|\mathrm{x_i}\rangle = \mathrm{x_i}|\mathrm{x_i}\rangle\,, \tag{II.5.1}$$

and the state vector would be written as $|\psi\rangle = \sum_\mathrm{i} \alpha_\mathrm{i}|\mathrm{x_i}\rangle$. The natural generalization for the continuous space case to write the following expression for the quantum state of a particle:

$$|\psi\rangle = \int \psi(\mathrm{x})|\mathrm{x}\rangle\ \mathrm{d}\mathrm{x}\,. \tag{II.5.2}$$

All we know about the particle state $|\psi\rangle$ is that the state is encoded in the corresponding complex function ψ of the continuous position variable x. The *sum* over the discrete subscript i gets replaced by *integral* over the continuous variable x, which is symbolically written as $\int\cdots\mathrm{d}\mathrm{x}$.

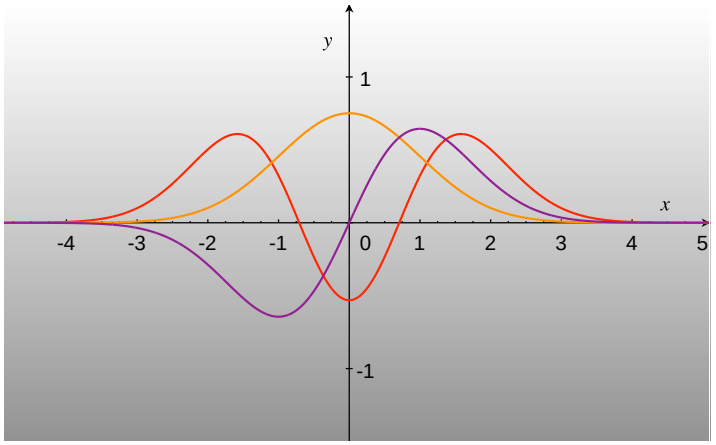


Figure II.5.4: *Harmonic oscillator wavefunctions.* wavefunctions of the three lowest energy states $\psi_n(x)$ with $n = 0, 1, 2$, of a quantum oscillator. The label n also gives the number of nodes: the even n functions are symmetric the odd ones are odd under $x \to -x$.

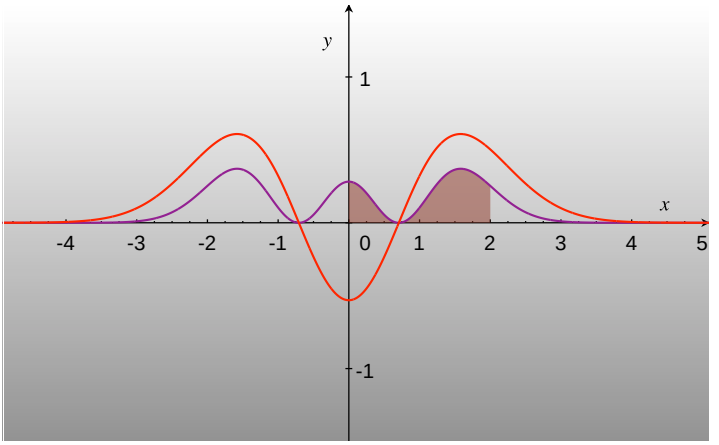


Figure II.5.5: *Harmonic oscillator probabilities.* The $n = 2$ wavefunction (red) and probability density (purple). The probability P_{02} is given by the shaded area. We talk about these states in detail in a later section on page 395.

And indeed, $\psi(x)$ is just the famous *wavefunction* that appears in the well-known Schrödinger equation we will get to later. To give you an idea we have depicted the three lowest energy states of a particle in a harmonic oscillator potential in Figure II.5.4. These will be discussed in more detail shortly. Talking heuristically one may say that the wavefunction represents nothing less than a vector in an infinite-dimensional vector space. In fact $\psi(x)$ is the '$|x\rangle$ component' of the state vector $|\psi\rangle$ which suggests that we should write it as such:

$$\psi(x) = \langle x|\psi\rangle, \tag{II.5.3}$$

leading to the expansion of the wavefunction in 'position eigen states',

$$|\psi\rangle = \int |x\rangle\langle x|\psi\rangle \, dx\,.$$

We have to make sure that we impose the *normalization condition* just as we did in the discrete case, in strict analogy it reads:

$$\langle\Psi|\Psi\rangle = \int \psi(x)^*\psi(x) \, dx = \int |\psi(x)|^2 \, dx = 1\,. \tag{II.5.4}$$

So what we learn is that quantum states of particles defined on a configuration space $\mathcal{X}$ correspond to elements of the space $\mathcal{H}$ of (complex) functions on $\mathcal{X}$ which are 'square integrable', meaning that they have to satisfy the condition (II.5.4). This space of square integrable functions is called the *Hilbert space*. One can also define a scalar product on the states that – not surprisingly – takes the form:

$$\langle\phi|\psi\rangle = \int \phi^*(x)\psi(x) \, dx\,,$$

completely analogous to formula (II.1.4). This once more underscores the exceptional elegance of Dirac's *bra* and *ket* notation.

You could say that by going from classical to quantum description we transcend from some space of coordinates to the space of functions on that space of coordinates. The difference with the description of the classical state is rather dramatic indeed, and you may wonder how to make sense out of it. What is the link of the wavefunction which is defined over all of space and the ordinary point-like particles we observe?

Probability interpretation. The interpretation is also completely in line with what we expect from the discrete case: $\psi(x)$ is the (complex) *probability amplitude* for the probability $p(x)$ of finding the particle at point x. The absolute square of the amplitude $p(x) = |\psi(x)|^2$ defines a *probability density*, and hence the probability P_{ab} of finding the particle in the range $a \leqslant x \leqslant b$ can be expressed as:

$$P_{ab} = \int_a^b p(x)\, dx\,. \qquad \text{(II.5.5)}$$

Formulas like the ones we have displayed in this section may at first look a bit daunting, and you may ask what the hell they mean. Well stay tuned in because it is not hard to visualize at all; the probability P_{ab} is just the area under $p(x)$ if you plot it as a function of x, between the points $x = a$ and $x = b$; This is depicted in Figure II.5.5 and for more details we refer to the *Mathematical Excursion on functions* in Appendix A of Part III.

As a matter of fact physicists love the bra and ket notation, it is compact and convenient to work with and it also keeps the conceptual structure of expressions remarkably transparent. And often progress originates in designing an optimal symbolic representation and notation.

This for the moment concludes our description of the space of quantum states that corresponds to a classical system with a continuous configuration space such as a particle moving in ordinary space. We saw that it is described by a complex wavefunction that may be considered as the components of a vector in an infinite-dimensional vector space of normalizable vectors which is called the Hilbert space. And we have mentioned that the square of the wavefunction corresponds to a probability density for where the particle can be found.

There are other pressing questions that immediately come to mind. You may ask: where did the velocity of the particle go, it appears nowhere in the specification of the quantum state? And what about its energy? Your point is well taken indeed – thank you – and we will return to the question of how, and to what extent, a precise velocity or momentum or energy can be assigned to a particle in the next section. But before we do so, I want to discuss an explicit example of a set of wavefunctions for a particle that lives not only in one dimension, but on a circle, which is a finite one-dimensional space without boundary.

A particle on a circle

In this subsection we turn to a concrete example and look at a quantum particle that lives on a unit circle with an angular coordinate $0 \leq \varphi \leq 2\pi$. This may strike you as a particularly useless theoretical problem, but one should be careful with those judgements. A lot of applications of physics and in particular quantum physics have to do with settings that are effectively low dimensional. Quantum wires are one-dimensional. A particle that is confined to the edge of a planar disc lives on a circle. In fact the groundbreaking Bohr-model of the atom amounted exactly to quantizing a particle on a circle, as he basically quantized the particle on classical circular orbits. Another example are 'quantum dots', which are basically finite two-dimensional domains on which particles can live.

A particle on a circle will be described by some complex wavefunction $\psi(\varphi)$ that is normalized but also has to satisfy a continuity or periodicity condition such that[3] $\psi(\varphi) = \psi(\varphi + 2\pi)$.

[3] It is more precise to say that this is a condition one imposes *a priori* on physical grounds. If there is some defect on the boundary one could well imagine to impose a different, non-trivial boundary condition, for example $\psi(\varphi + 2\pi) = e^{i\gamma}\psi(\varphi)$. A more sophisticated treatment of the problem would be to say that we extend the set of observables to arbitrary translations x and decompose these into $x = 2n\pi + \varphi$. The discrete translations by $2n\pi$ form an invariant subgroup Z of the group of translations on the real line $\mathbb{R}$; the different boundary conditions form representations of this Z group and these are labeled by the angle $0 \leq \gamma < 2\pi$.

Figure II.5.6: *La Danse.* Circle dance by the French painter Henri Matisse, painted in 1910. (©Succession Henri Matisse.)

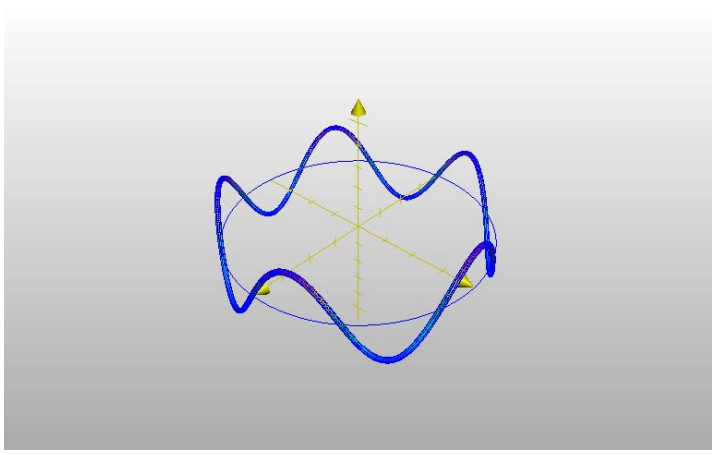


Figure II.5.7: *Going in circles.* We have depicted the $k = 5$ real wavefunction $\psi(\varphi) = \frac{1}{\sqrt{2\pi}}\cos(5\varphi)$ of a particle on a circle,

Momentum eigenstates. The periodic solutions are of the form

$$\langle \varphi | k \rangle = \psi_k(\varphi) = \sqrt{\frac{1}{2\pi\hbar}} e^{ik\varphi} . \qquad \text{(II.5.6)}$$

You would expect maybe periodic functions like cosines and sines, but as we allow complex functions it is much more natural to write then as simple exponential functions, and in a sense it amounts to the same thing because of that beautiful Euler identity $e^{ik\varphi} = \cos(k\varphi) + i\sin(k\varphi)$ as is explained in the *Math Excursion* about complex numbers on page 607 of Part III. The periodicity condition leads to the condition that $e^{2\pi ik} = 1$, which is satisfied only if k is restricted to integer values.

Observe that these periodic states $\psi_k(\varphi)$ have a wavelength $\lambda = 2\pi/k$ and using the relation of De Broglie $\lambda = h/p$ says that in these particular periodic states the particle carries a momentum $p_k = \hbar k$. What about the energy of the particle? If we think of a free particle with no force on it, the energy would just be the kinetic energy of the states $E_k = p_k^2/2m$ and therefore grows proportional to k^2. At this point, however, we could also assume that the particle is a relativistic particle, in which case the expression for the energy of the k-th mode would be $E_k = \sqrt{p_k^2 c^2 + m^2 c^4}$ which for small momentum reduces to the previous expression but for large p_k we would get $E_k \simeq p_k c$ which is proportional to k. We can also immediately calculate the probability distribution for the states to equal:

$$p_k(\varphi) = \psi_k^*(\varphi)\psi_k(\varphi) = \psi_{-k}(\varphi)\psi_k(\varphi) = \frac{1}{2\pi} .$$

This probability density for where to find the particle is constant! This tells us that whereas the momentum of the particle is completely fixed with zero uncertainty, the position of the particle is maximally uncertain, because it corresponds to a uniform distribution over all of space. In these states there is no preference whatsoever for any position or region. The conclusion is that in this momentum framework the particle logically speaking has no position. A dramatic instance of the Heisenberg uncertainty principle. We return to this point later on, when we will show wavefunctions which to a certain extent look more like localized particles. These wavefunctions will be particular linear superpositions of this set of momentum eigenstates.

Just like in the discrete case for a general state we may

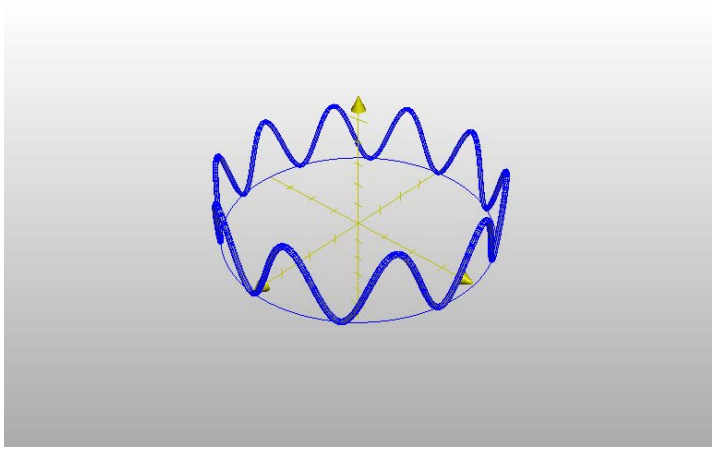


Figure II.5.8: *Going in circles.* The probability distribution $\rho(\varphi) = \frac{1}{2\pi}\cos^2(5\varphi)$ of a particle on a circle.

write a general expansion like:

$$|\psi\rangle = \sum_k \alpha_k |k>, \tag{II.5.7}$$

where the basis states $|k\rangle$ would equal

$$|k\rangle = \int \langle\varphi|k\rangle\ |\varphi\rangle\ d\varphi = \sqrt{\frac{1}{2\pi}} \int e^{ik\varphi}\ |\varphi\rangle\ d\varphi\,.$$

These states form an orthonormal basis, meaning that they satisfy the orthonormality condition[4]

$$\langle k|k'\rangle = \int \langle k|\varphi\rangle\langle\varphi|k'\rangle d\varphi = \delta_{kk'}\,.$$

Now the sum in. the expression (II.5.7) extends over all integer values of k, indeed confirming our expectation that the quantum state of a particle on a circle is like a vector in an infinite-dimensional space. There is also a completeness relation for the basis in analogy with equation (II.2.10) which reads

$$\sum_k |k\rangle\langle k| = 1\,.$$

[4] To prove it you need the functional relation that $\frac{1}{2\pi}\int \exp(i(k-k')\varphi)\ d\varphi = \delta_{kk'}$.

Position and momentum operators

In the earlier chapters we have been talking about quantum dynamical variables as operators or matrices. For example for the qubit we showed in Chapter II.1 that we could interpret the Pauli Z-matrix as the position operator, and the Pauli X-matrix as the momentum operator. So the first question that comes up if we think of particle states as wavefunctions, what the operator valued observables should look like. Something like infinite-dimensional matrices maybe? The answer is simpler than that and quite natural if you think of operators that have to act on functions. You can multiply functions by other functions, but more importantly we can differentiate functions. We should expect dynamical variables to be represented by differential operators. So let us first consider the momentum operator.

Momentum. In this section we look at the definition of momentum and position operators for a particle on a circle. The state vectors are the wavefunctions $\psi_k(\varphi)$ given in equation (II.5.6), we will show that the momentum corresponds to the differential operator,

$$P = -i\hbar\frac{d}{d\varphi}\,.$$

First observe that the functions $\psi_k(x)$ are eigenfunctions of P, because,

$$P\psi_k(\varphi) = \hbar k\psi_k(\varphi)\,,$$

and recall that we argued in the previous chapter based on De Broglies heuristic argument that the momentum of a particle in the k-th state is indeed equal to $p_k = \hbar k$.

Generator of translations. At this point you might want to look at the *Math Excursion* on page 607 of Part III, where it is shown that the displacement of the state vector or wavefunction is also generated by the derivative or differential

operator[5] $K = \frac{d}{d\varphi}$. We can formally put it in the exponent, just like the sigma matrices before, having the property:

$$e^{i\theta K}\psi_k(\varphi) = e^{i\theta k}\psi_k(\varphi) = \psi_k(\varphi+\theta)\,. \qquad \text{(II.5.8)}$$

We see that this exponential operator just shifts the argument of the function by the factor in front of K in the exponent. This equation is the precise mathematical expression of the statement that the momentum operator (in fact $P/\hbar$ to be precise) acting on a function 'generates' spatial translations of its coordinate (position).

The position operator. What about the position operator Φ? It acts on the wavefunction as $\Phi\ \psi(\varphi) = \varphi\psi(\varphi)$, i.e. by just multiplying the wavefunction by the *variable* 'φ'. Note, however, that the $\psi_k(\varphi)$ are eigenfunctions of momentum P but *not* of position Φ (because k is a constant and φ is not, it is a coordinate, a variable). So the position operator 'multiplies' the wavefunction with the *function* 'φ'.

It may be useful to again point out the analogy with the qubit case in Chapter II.2, where the would-be position operator was Z and a would-be momentum operator could be X. We could then consider the states $|\pm\rangle$ defined in equation (II.2.4), which are eigenstates *not* of Z but of X, because $X|\pm\rangle = \pm\ |\pm\rangle$. And indeed, acting with Z on, for example the X eigenvector $|+\rangle$ would multiply each component with a different coordinate value, leading to $Z|+\rangle = |-\rangle$. So, acting with the coordinate operator on a momentum eigenstate changes it to another state.

Canonical commutation relations. Being eigenfunctions of momentum, the $\psi_k(\varphi)$ are also eigenfunctions of a Hamiltonian $H = P^2/2m$ describing a free particle of mass m that moves on a circle. We also see that as we might expect the momentum and position operators do not commute, they satisfy the so-called *canonical commutation relations*:

$$[X, P] = i\hbar\,. \qquad \text{(II.5.9)}$$

To see that this is true it is most convenient to think of the commutator as an operator working on a (wave) function $f(x)$, then we obtain:

$$\begin{aligned}[X,P]\ f(x) &= -Xi\hbar\frac{df(x)}{dx} + i\hbar\frac{d}{dx}\ X\ f(x)\\ &= -i\hbar x\frac{df(x)}{dx} + i\hbar\frac{d}{dx}\ (x\ f(x))\\ &= i\hbar(\frac{d}{dx}\ x)f(x) = i\hbar f(x)\,.\end{aligned}$$

As the function appearing on both sides of the equation is *arbitrary* we may conclude that the statement (II.5.9) is true as a property of the operators.

Raising and lowering. Let us ask for the *raising* and *lowering* operators of this problem. Let us first try to find operators $Q_\pm$ that satisfy the commutation relations:

$$[P, Q_\pm(X)] = \pm a Q_\pm(X)\,, \qquad \text{(II.5.10)}$$

and as

$$PQ_\pm - Q_\pm P = -i\hbar\frac{dQ_\pm}{dx}\,,$$

we obtain an equation for the functions $Q_\pm(x)$:

$$-i\hbar\frac{dQ_\pm(x)}{dx} = \pm a Q_\pm(x)\,.$$

The solutions to this equation are $Q_\pm(x) = c\exp(\pm iax/\hbar)$, and therefore one obtains for the operators:

$$Q_\pm(X) = c\ e^{\pm iaX/\hbar}\,. \qquad \text{(II.5.11)}$$

The interpretation is now as follows. The momentum of a particle on the circle has a discrete spectrum $\{\hbar k\}$ with integers $\infty < k < \infty$, for clockwise and counterclockwise moving particles. The smallest possible momentum state

[5] The difference between P and K is a matter of units or dimensions, the dimension of the differential operator is $[1/\mathrm{length}]$ to get the dimensions of momentum we have to multiply by a constant with dimension $[\mathrm{lenght}\times\mathrm{momentum}] = [\mathrm{joule}\times\mathrm{second}]$ and yes – not surprising – that constant is nothing but Planck constant $\hbar$.

has $k = 0$ and the raising and lowering operators (II.5.11) sequentially generate all the eigenfunctions $\psi_k(x)$ if we choose $a = 1$. We clearly have to adjust the value of a to comply with the imposed boundary condition.

Heisenberg's uncertainty. It is amusing to check the Heisenberg uncertainty relation by verifying that indeed $\Delta x = L$ and $\Delta p = p_0 = \hbar\pi/L$ satisfy:

$$\begin{aligned}\Delta x \Delta p &= \hbar\pi \geq |i\langle\psi_0|[X, P]|\psi_0\rangle| \\ &= \hbar\langle\psi_0|\psi_0\rangle = \hbar\,.\end{aligned}$$

We see that these states do not saturate the lower bound on the uncertainty relation. That lower bound is $\hbar/2$, as we showed in Chapter II.2 on page 317.

Energy generates time evolution

Time evolution. If we talk about time evolution of a classical system, we think in the first place of Newton, but in the realm of computation we also think of the physical implementation of a sequence of logical gates. A computation is in that sense a discrete dynamical process whose rate is set by the speed or clock time of the chip, today being of the order of nanoseconds. We process information by manipulating it through interacting with it in a controlled way by having logical gates acting. That is similar to applying a force to get ourselves moving as we saw in the previous chapters. Now even in the heyday of classical mechanics many different approaches were formulated in attempts to solve specific dynamical problems by people like Hamilton, Jacobi, Laplace, Lagrange, Legendre and others. We discussed some of them in Chapter I.1 on page 16.

In Figure II.5.9 we have indicated various paths that lead from the domain of classical mechanics to the corresponding quantum equations. I am going to discuss them sequentially, and start with the Schrödinger equation.

Wave mechanics: the Schrödinger equation

The wavefunction of a quantum system evolves in time according to the famous Schrödinger equation. Dynamical changes in a physical system are induced by the underlying forces acting on the system and between its constituent parts, and their effect can be represented in terms of what is called the energy or Hamiltonian operator H. For a single qubit system the operators can be represented as 2×2 matrices, for a two-qubit system they are 4×4 matrices, etc. The Schrödinger equation can be written

$$i\hbar\frac{d|\psi(t)\rangle}{dt} = H|\psi(t)\rangle\,. \tag{II.5.12}$$

This is a linear differential equation expressing the property that the time evolution of a quantum system is generated by its energy operator. Assuming that H is constant, given an initial state $|\psi(0)\rangle$ the solution is simply

$$|\psi(t)\rangle = U(t)|\psi(0)\rangle \text{ with } U(t) = e^{-iHt/\hbar}\,. \tag{II.5.13}$$

The time evolution is *unitary*, meaning that the operator $U(t)$ satisfies $UU^\dagger = 1$.

$$\begin{aligned}U^\dagger &= \exp(-iHt/\hbar)^\dagger \\ &= \exp(iH^\dagger t/\hbar) = \exp(iHt/\hbar) = U^{-1}\,.\end{aligned} \tag{II.5.14}$$

Unitary time evolution means that the length of the state vector remains invariant, which is necessary to preserve the total probability for the system to be in any of its possible states. The unitary nature of the time evolution operator U follows directly from the fact that H is hermitian: $H^\dagger = H$. Any hermitian 2×2 matrix can be written

$$A = \begin{pmatrix} a & b + ic \\ b - ic & -a \end{pmatrix},$$

where a, b and c are real numbers.[6]

[6] We omitted a component proportional to the unit matrix as it acts trivially on any state. We speak of the part that has no trace.

Stationary states. From equation (II.5.13) results it is also immediately clear what the importance is of the eigenstates of the Hamiltonian – the energy eigenstates. An energy eigenstate $|\psi_n\rangle$ satisfies by definition $H|\psi_n\rangle = E_n|\psi_n\rangle$, and thus for such states:

$$|\psi(t)\rangle = \exp(iE_n t/\hbar)\,\psi(0)\rangle\,. \qquad \text{(II.5.15)}$$

The state is not quite time-independent, but it changes only by an overall phase factor, which means that the probability density $|\psi|^2$, or the expectation value of any operator will not change over time. The state is strictly speaking not *static* and therefore called *stationary*.

Time dependence. But if we act on a state that is not an eigenstate of the energy, we get a time dependent solution. For the simple example of a single qubit, suppose the initial state is

$$|\psi(0)\rangle = |+\rangle \simeq \sqrt{\frac{1}{2}}\begin{pmatrix}1\\1\end{pmatrix}\,.$$

On the right, for the sake of convenience, we have written the state as a column vector. Consider the energy of a spin in an external magnetic field B directed along the positive z-axis.[7] In this case H is given by $H = bZ$. Now the initial state is a linear combination of two different energy eigenstates. From equation (II.5.13) it follows that,

$$\begin{aligned} U(t) &= \exp(\frac{-ibt}{2\hbar}Z) \\ &= \begin{pmatrix}\exp(-ibt/2\hbar) & 0\\ 0 & \exp(ibt/2\hbar)\end{pmatrix}\,. \end{aligned} \qquad \text{(II.5.16)}$$

We obtain an oscillatory time dependence for the state, not just a phase factor, i.e.

$$\begin{aligned} |\psi(t)\rangle &= \sqrt{\frac{1}{2}}\begin{pmatrix}e^{-ibt/\hbar}\\ e^{ibt/\hbar}\end{pmatrix} \\ &= \sqrt{\frac{1}{2}}\left[\cos\frac{bt}{\hbar}\begin{pmatrix}1\\1\end{pmatrix} + i\sin\frac{bt}{\hbar}\begin{pmatrix}-1\\1\end{pmatrix}\right]\,. \end{aligned} \qquad \text{(II.5.17)}$$

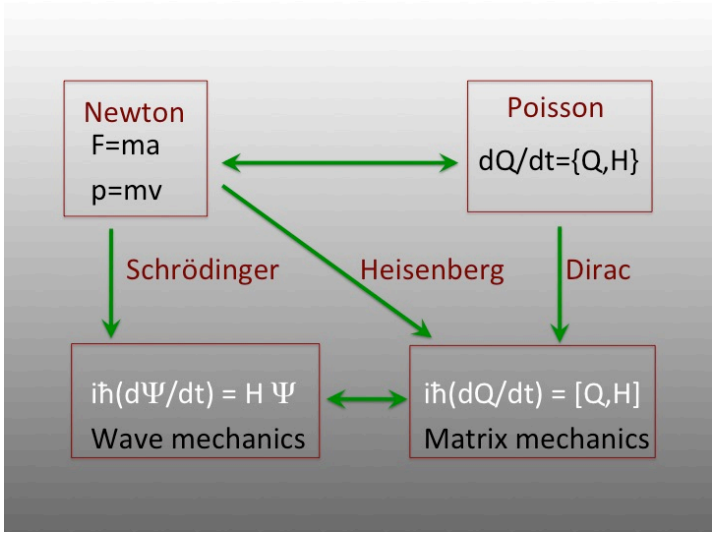


Figure II.5.9: *Ways to go quantum.* Various pathways from different but equivalent formulations of classical mechanics to the Schrödinger and Heisenberg – also equivalent – formulations of quantum theory.

The state oscillates between the $|+\rangle$ with probability $p_+ = |\langle+|\psi(t)\rangle|^2 = \cos^2 bt/\hbar$ and $|-\rangle$ with $p_- = |\langle-1|\psi(t)\rangle|^2 = \sin^2 bt/\hbar$. This simple example applies in some form or another to numerous physically relevant two-level systems.

We see that, in contrast to classical mechanics, the time evolution equation is first order in time and linear in the wavefunction. In general the Hamiltonian can be a complicated function of the basic dynamical variables and therefore it is only in rare situations that one can find an exact analytic solution. On the other hand it is also surprising to see how a relatively small number of exactly solved problems can serve to get a deep insight in, and feeling for, what kind of behavior quantum systems exhibit.

[7] Quantum spins necessarily have a magnetic moment, so in addition to carrying an intrinsic angular momentum they also interact with a magnetic field.

Matrix mechanics: the Heisenberg equation

In the previous section we have considered the time evolution of the state generated by some particular Hamiltonian which we assumed to be time independent. In that Schrödinger type description the operators one considers are mostly time independent and the time-dependent state $|\psi(t)\rangle$ is a solution to the Schrödinger equation with the given Hamiltonian which characterized the system and its interactions. There is a complementary view which was developed by Heisenberg, it is often called *'matrix mechanics'* which lead to the Heisenberg equation. In his view, which in a sense is closer to classical dynamics, the dynamical variables, meaning observables like matrices, are the objects that change in time whereas the state remains fixed. The simplest way to see how this comes about is to rewrite the definition of the expectation value of an operator A in a suggestive way as:

$$\begin{aligned}\langle\psi(t)|A|\psi(t)\rangle &= \langle\psi(0)|e^{iHt/\hbar}Ae^{-iHt/\hbar}|\psi(0)\rangle \\ &= \langle\psi(0)|A(t)|\psi(0)\rangle\,. \qquad \text{(II.5.18)}\end{aligned}$$

In other words we have defined time-dependent observables for the system through the relation:

$$A(t) \equiv e^{+iHt/\hbar}Ae^{-iHt/\hbar}\,.$$

By calculating the time derivative of the above expression one arrives at Heisenberg's quantum equation of motion:

$$i\hbar\frac{dA(t)}{dt} = [H, A(t)]\,, \qquad \text{(II.5.19)}$$

and we have an equation that tells us that the time evolution of operators acting on the state space is generated by the commutator with the Hamiltonian of the system. Note that the commutation relations of observables are unchanged by the transformation, so we still have the canonical commutator $[X, P] = i\hbar$. I would also like to remind the readers who happened to read my discussion on Poisson brackets on page 16 of Part I, that there is indeed a striking similarity between the classical Poisson brackets and the Heisenberg commutator equations. The recipe is to make in equations (I.1.14) to (I.1.16) the following replacement

$$\{\ ,\ \}_{pb} \ \Rightarrow\ -\frac{i}{\hbar}[\ ,\ ],$$

to obtain the canonical quantum equations! This was by the way the method Dirac used to 'quantize' systems.

It is illuminating to keep both formulations in mind. Certain questions can be answered more easily in the Schrödinger picture and others in the Heisenberg picture.

Symmetries and conservation laws. The Heisenberg equation yields a direct understanding of the existence of 'constants of the motion' or conservation laws. For physical variables described by operators Q that commute with the Hamiltonian i.e. $[H, Q] = 0$, the Heisenberg equation teaches us that $dQ/dt = 0$ and thus that Q is conserved in time. Such operators that commute with the Hamiltonian are by definition called *symmetry operators*. You see that energy is one of them, and that had better be so, because time independence of the Hamiltonian was after all our starting point. Depending on the system and its Hamiltonian we will find out about the conserved quantities this way, like momentum, angular momentum, the Lenz vector, charge, isospin etc. Indeed summing up these examples one realizes how important these basic conservation laws are, as they allow us to characterize the states by properties that are robust in time, and that allows us to label and assign names to things! After all, your name would be useless if it were to change every day.

Degeneracies. The other consequence of having conserved quantities is that if Q acts on an eigenstate of the Hamiltonian then it may well make another state, but that state will have the same energy as the first one. You can use the conserved quantities or symmetry operators to generate 'degenerate states' in the spectrum. The statement is stronger than that, because you can always find enough symmetry operators to resolve all degeneracies

and label the different orthogonal states that are degenerate in energy by labels referring to conserved quantum numbers. We saw this principle already at work in Chapter I.4 where we discussed the discovery of the electron spin. This was achieved by lifting the degeneracy by introducing an external magnetic field, which broke the rotational symmetry of the system.

A framework of symmetry operators. So in choosing a framework, we often like to include the Hamiltonian as one of the operators. The next thing you may want to do is to add operators that commute with H, which in other words correspond to conserved quantities. These operators form a closed algebra in the sense that if Q_1 and Q_2 commute with H, then also their commutator $[Q_1, Q_2]$ will commute with H. This way we can construct a commutator or Lie-algebra of symmetry operators including the Hamiltonian. This algebra is then called the *symmetry algebra* for the system.

Next we follow the instructions for a consistent framework and select, out of all those conserved quantities, a maximal number which do *mutually* commute. That defines a sub-algebra of the full symmetry algebra, consisting of observables whose combined eigenvalues form the sample space, for that framework. In fact such a maximal set of mutually commuting independent symmetry operators is called the *Cartan subalgebra* of the symmetry algebra. This algebra is named after the famous French mathematician Élie Cartan, who succeeded in completely classifying all possible finite-dimensional (complex) Lie-algebra's. Many of those play a crucial role in quantum physics.

The next chapter is devoted to different kinds of symmetry and their breaking, and it will become clear that the notion of symmetry is one of the guiding principles that has played a leading role in the development of modern physics.

Generators of symmetries. So we have arrived at a rather quantessential picture linked to (continuous) symmetries. The operators Q that are conserved generate the symmetries, and they can therefore be used to label the states, and furthermore they are physical observables. If I say that the Heisenberg equation just tells you that the Hamiltonian generates a time translation, what I mean is that an infinitesimal change of an observable A in time, $-i\hbar\, dA/dt$ equals the commutator with the Hamiltonian. One can also write for example:

$$-i\hbar\frac{dA}{dx} = [P, A]\,,$$
$$i\hbar\frac{dA}{dp} = [X, A]\,,$$

which states that the operator dependence on position or momentum is generated by their 'duals,' the momentum and position operators respectively. Similarly the commutator with the angular momentum component L_z generates an infinitesimal rotation around the z-axis of the operators. We see that the Heisenberg equation is in fact one of many. It is the equation that expresses the time-translation-symmetry of the underlying space-time, from which energy conservation is derived.

Classical lookalikes

In our discussion of (free) particle states we have clearly found two extremes: (i) the momentum states $|k\rangle$, where the momentum and energy have no uncertainty, but the uncertainty in position is maximal and (ii) the position states $|\varphi\rangle$, where the converse would hold. Neither of these seems close to what we think about when we talk about a classical particle moving on a circle. We know that we are free to consider any state of the type given in (II.5.2) and therefore we can ask whether it is possible to find a particular linear combinations of basic quantum states that look more like the classical ones.

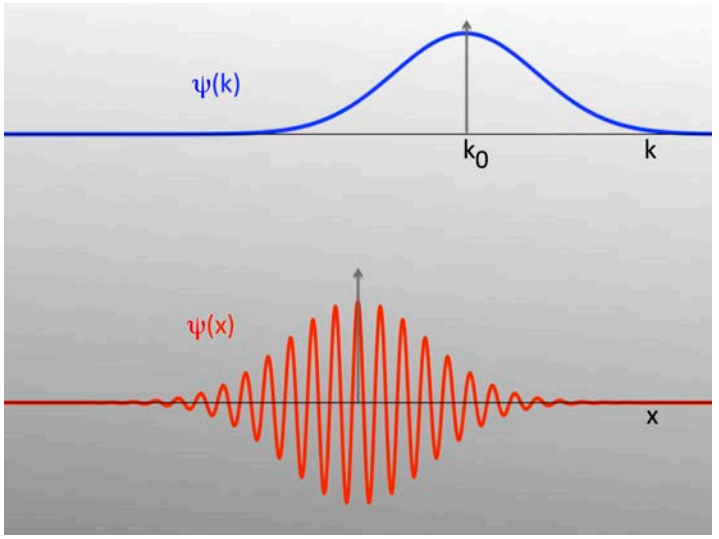


Figure II.5.10: *Wave packets.* We make a 'gaussian' superposition of plane waves with momentum k given by $\psi(k)$ (the blue curve). Then the wavefunction in x-space (the red curve) is as depicted at the bottom, and the enveloping curve of the wavy pattern is again a gaussian.

Wave packets. This is certainly possible, as actually Schrödinger already pointed out. He studied what are called 'wave packets.' These are smooth linear combinations of say momentum eigenstates, which are localized in both position and momentum space. These packets have an average momentum k_0 and an average position and look in many aspects as extended particle-like objects.

The starting point is simple, namely to look for states where the uncertainty in canonically conjugate (incompatible) variables is minimized and balanced, respecting the Heisenberg uncertainty relations. But because the Schrödinger equation is linear we may consider arbitrary linear combinations of the states, and are then time-dependent solutions because the different momentum components have different energies.

Such a wave packet can be defined by specifying a function $\psi(k)$ and looking at the state

$$|\psi\rangle = \int \psi(k)|k\rangle\, dk\,.$$

As the formula suggests the function is just the ' wavefunction in momentum space', as we may write:

$$\psi(k) = \langle k|\psi\rangle\,.$$

Now let us take a smooth gaussian (normal distribution) centered around some momentum k_0,[8]

$$\psi(k) = (\frac{2\alpha}{\pi})^{1/4} e^{-\alpha(k-k_0)^2}\,.$$

The factor in front makes sure that the state is properly normalized, so that all probabilities add up to one. We have displayed this function in Figure II.5.10 and indeed it is nicely peaked with a certain width around k_0.

Now we want to see what this package deal means for people who live in ordinary x or φ space. Using equation (II.5.6) we calculate:

$$\begin{aligned}\psi(\varphi) &= \langle\varphi|\psi\rangle = \int\langle\varphi|k\rangle\langle k|\psi\rangle\, dk = \\ &= (\frac{1}{2\pi\alpha})^{1/4} e^{ik_0\varphi} e^{-\hbar\varphi^2/4\alpha}\,.\end{aligned}$$

What do we see? First of all we see that the wavefunction of the state is also gaussian in φ space! That is nice because it does indeed mean that the packet is also well localized in position space, just as we wanted it. What we also see is that the width of that distribution is like the inverse of the width in momentum space. To be precise we have $\Delta k = \sqrt{1/2\alpha}$, and $\Delta\varphi = \sqrt{\alpha/2}$, which shows that the packet is optimal in the sense that it saturates the lower bound on the uncertainties imposed by Heisenberg: $\Delta k\ \Delta\varphi = 1/2$. Finally we see that the wavefunction in position space also has a factor $e^{ik_0\varphi}$, which makes the

[8] We discussed the gaussian or normal distribution in the *Math Excursion* on probability and statistics at the end of Chapter I.1. There it was also explained why this distribution pops up everywhere.

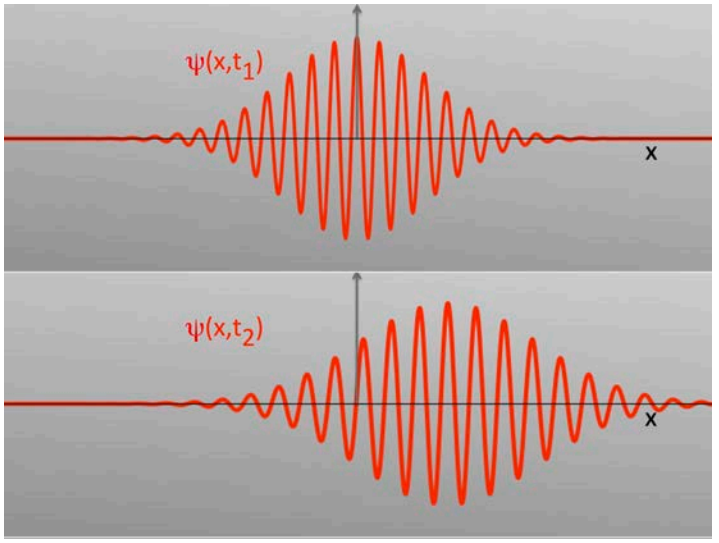


Figure II.5.11: *Wave packet dispersion.* The time evolution of the free wave packet according to the Schrödinger equation has two generic features:(i) it moves forward with the group velocity of the package which is the effective velocity of the 'particle', and (ii) the package will broaden (disperse) over time.

function periodic (and complex). The red curve depicted in the figure is (the real part of) the wave packet in coordinate space.

Propagation and dispersion. Let us assume that the configuration at $t = 0$ is the one we just described, then it is interesting to see what happens in time exactly because the packet is made up of momentum components that propagate with different speeds. The question is therefore what that means for the time evolution of the packet as a whole. We don't want to go through the calculation here, but the important message is sketched in Figure II.5.11. The first point to mention is that the center of the packet, or the *envelope* of the wavy pattern, moves with the so-called *group velocity*. This is a velocity which is different from the *phase velocity* of the individual momentum components. One typically sees the effect that the wavy pattern moves faster than the envelope and one sees the small wave appear on the left (increasing of amplitude) and disappear at the right (decreasing of amplitude) of the envelope. The second point to mention is that the packet broadens in time, it *disperses*. If one calculates the probability distribution $p(\varphi)$ the periodicity drops out and we get a pure gaussian that is broadening, as we displayed already in Figure II.5.3 for a particle at rest. This dispersion worried among others Schrödinger himself quite a bit, because it basically blocked a direct interpretation of the wave packet as a particle, which basically seemed to disintegrate on quite short time scales. It was this aspect that was resolved by the probabilistic interpretation (the so-called *Copenhagener Deutung*) proposed by Born.

Raising and lowering operators. 

Let us briefly talk about yet another way to represent the general particle state (II.5.7), which utilizes ladder or raising and lowering operators that are completely analogous to what we did for the qubit in (II.2.13) and (II.2.14). First we look for an operator that can step from a state $|k\rangle$ to $|(k+1)\rangle$. Consider the following so-called step operators:

$$t_\pm = e^{\pm i\Phi} \tag{II.5.20}$$

where Φ is the coordinate operator given in (II.5.1) that satisfies $\Phi\ |\varphi_0\rangle = \varphi_0\,|\varphi_0\rangle$. Applying $t_\pm$ yields

$$\begin{aligned} t_\pm|k\rangle &= e^{\pm i\Phi}\,\sqrt{\frac{1}{2\pi}}\int e^{ik\varphi}\,|\varphi\rangle\,d\varphi \\ &= \sqrt{\frac{1}{2\pi}}\int e^{ik\varphi}\,e^{\pm i\varphi}\,|\varphi\rangle\,d\varphi = |k\pm 1\rangle \end{aligned} \tag{II.5.21}$$

where we let the operators act on the state $|\varphi\rangle$ in going from the first to the second line. So with $t_\pm$ one may step through the spectrum.

This is not yet what we want; what we really want is operators that start from some lowest energy states. The energy of the free particle state to be equal $E_k = p_k^2/2m = \hbar^2k^2/2m$ then the lowest energy state is $|0\rangle$ with $E_0 = 0$. We like the right and left moving states to be generated from some lowest energy states. Consider then, instead of

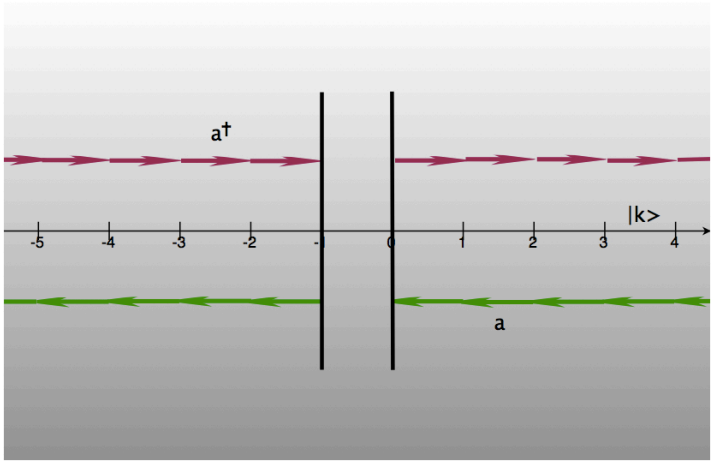


Figure II.5.12: *Step operators.* Action of the ladder or step operators $a^\dagger$ (raising-purple) and a (lowering-green) defined in equation (II.5.22) on the space of states labeled by $|k\rangle$. There are two sectors: the right movers with $k \geq 0$ and the left movers with $k \leq -1$.

the operators $t_\pm$, the two related ladder operators[9]:

$$a = e^{-i\Phi}P, \quad a^\dagger = Pe^{i\Phi}, \tag{II.5.22}$$

these satisfy interesting commutation relations:

$$[a, a] = [a^\dagger, a^\dagger] = 0, \quad \text{and} \quad [a, a^\dagger] = 2P + 1. \tag{II.5.23}$$

Furthermore we see that for a free particle (with m=1) we can write the Hamiltonian as:

$$H = \frac{1}{2}a^\dagger a = \frac{1}{2}P^2. \tag{II.5.24}$$

If we apply the operators to some state $|k\rangle$, we obtain:

$$a^\dagger\,|k\rangle = (k+1)|k+1\rangle\,,\ a|k\rangle = k|k-1\rangle\,,$$

which illustrates the fact that these operators basically raise or lower the momentum of the state by one unit. These constructions demonstrate two surprising properties of the states and operators. With these operators you can indeed walk through the sample space of states but you will run into certain 'no trespassing' signs , where the next step you make you would let you disappear into nothing! The first one tells you that if you act with a you may come down from positive k al the way down to $k = 0$, but not any further because $a|0\rangle = 0$. However if you start from $|k = -1\rangle$, then a will walk you down all the way to minus infinity. Something similar happens with $a^\dagger$: it walks you up from any negative value until you hit $|-1\rangle$ where it halts, but starting at $|0\rangle$ it will bring you all the way to plus infinity. What this means that the spectrum naturally breaks up into two pieces: one of which you could define as the right movers with $k \geq 0$ and the other as the left movers with $k < 0$.

State operators. We can now also construct operators that directly create any momentum state from the ground-state. For example the state $|k\rangle$ can be obtained by acting k times with $a^\dagger$ on the ground state $|0\rangle$, as the following calculation shows:

$$|k\rangle = \frac{a^\dagger}{k}|k-1\rangle = \ldots = \frac{(a^\dagger)^k}{k!}|0\rangle\,.$$

The general state $|\Psi\rangle$ could also be symmetrically represented like an operator :

$$|\Psi\rangle = \Psi\,|0\rangle = (\alpha_0 + \Sigma_{k=1}\frac{1}{k!}(\alpha_k a^{\dagger k} + \alpha_{-k}b^k))\,|0\rangle\,, \tag{II.5.25}$$

where we have defined what you could call a 'particle-state' operator Ψ and b is a shifted operator $b = e^{-i\Phi}(P - 1)$. The correspondence between states and these type of step operators acting on a 'vacuum' or 'ground' state will be of great use if we move from quantum particles to quantum fields as we will do in the next section. ■

[9] In the remainder of this section we set $\hbar = 1$, to keep the formulas simple.

The harmonic oscillator

Oscillators everywhere! The harmonic oscillator or the 'particle in a harmonic oscillator potential' is a system that is treated extensively in any book on quantum physics and classical physics alike. In spite of the fact that we do not see swinging pendulums all over the place, the simple truth is that the world around us is actually largely made up of oscillators! One way to understand that is to realize that most 'things' are in a state of equilibrium, in other words they are in a state of minimum energy. And yes, if you perturb a system in equilibrium, it will start to oscillate about its equilibrium state. You knock on the table, you drop a stone in the lake, the days, the seasons, economic cycles, the orientation of the Earth's axis, the strings of your guitar and of string theory, the rhythms of life: all are oscillatory motions in some suitable space.

So imagine the horizontal axis describing the displacement of some relevant variable from equilibrium, and let us call that variable, yes indeed, x, then along the vertical axis we plot the energy V (as a function of x). This function generically will have a particular shape. It will have a minimum at $x = 0$, and if we think that we study small perturbations we might look at V(x) close to the origin and describe it effectively as an expansion in (positive) powers of x. The first term would be linear, but that could not be because it would not correspond to a minimum anymore, the minimum would have shifted away. So the first relevant term would be the quadratic term which we write as $V(x) = \frac{1}{2}\omega^2 x^2$. You get the bowl-shaped potential depicted in Figure II.5.14. In Newtonian mechanics this would imply a force $F = -dV/dx = -\omega^2 x$, thus a linear force trying to move the system back to the equilibrium position. This is not surprisingly called a harmonic force. As you can think of a marble rolling forth and back in the bowl. We discussed this dynamical system at length in Chapter I.1. It is important to now look at quantum oscillators because the microscopic world is also beset with them. This is a model

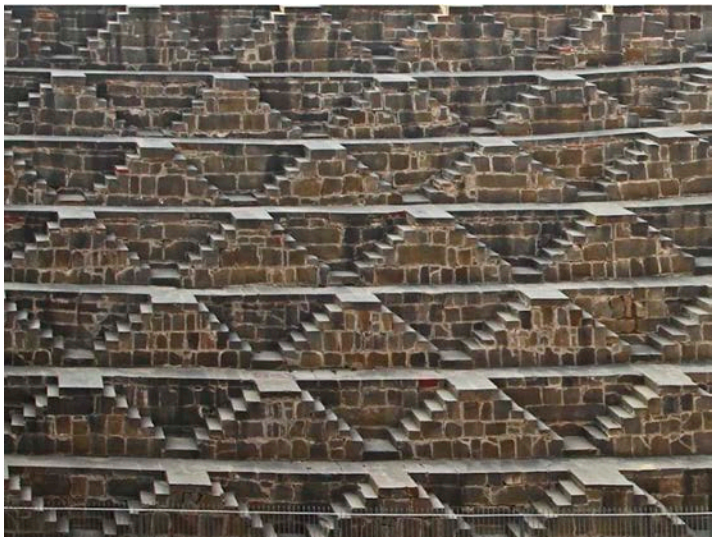


Figure II.5.13: *The stepwell of Chand Baori. These remarkable stepwells in India were once used to store water. Chand Baori is made up of 3.500 steps over 13 stories. The steps look like states forming a discrete spectrum of some quantum system.*(Source: Wikimedia.)

system that at first looks like one of these totally boring academic, dry-nerd-drill-home-trainer kind of things. The deadliest didactic horse ever. No! Imagine, its applications on all rungs of the quantum ladder are quite stunning and we will come across a few of them. So, please stay with me for this one.

If we return to basics, our starting point is the simple Hamiltonian for a unit mass particle in a harmonic potential:

$$H = \frac{1}{2}(p^2 + \omega^2 x^2)\,. \qquad \text{(II.5.26)}$$

The classical equations are,

$$\frac{dx}{dt} = p\,, \qquad \frac{dp}{dt} = -\omega^2 x\,.$$

We will treat these equations in the Heisenberg picture meaning that we have time dependent operators $X(t)$ and $P(t)$ with the canonical commutation relations:[10] $[X, P] =$

[10] if we postulate them at $t = 0$, the unitary time evolution ensures

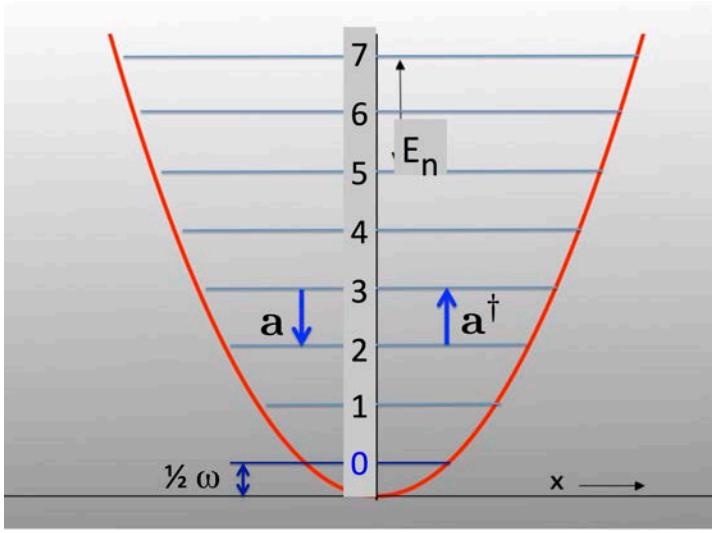


Figure II.5.14: *Harmonic oscillator.* Action of the ladder or step operators $a^\dagger$ (raising) and a (lowering) defined in equation (II.5.29) on the space of energy eigenstates $|n\rangle$. The ground state $|0\rangle$ has energy $E_0 = 1/2\omega$.

$[X(t), P(t)] = i\hbar$. Interestingly the equations can then be solved by using (commutator) algebra only. These are coupled equations, and you can decouple them by using the complex linear combinations which are as we will see *raising and lowering operators*:

$$a(t) = \sqrt{\frac{1}{2\omega}}(\omega X + iP) \equiv \tilde{X} + i\tilde{P} \qquad \text{(II.5.27)}$$

$$a^\dagger(t) = \sqrt{\frac{1}{2\omega}}(\omega X - iP) \equiv \tilde{X} - i\tilde{P}\,. \qquad \text{(II.5.28)}$$

The solutions have simple phases:

$$a(t) = ae^{-i\omega t} \qquad a^\dagger(t) = a^\dagger e^{+i\omega t}\,, \qquad \text{(II.5.29)}$$

these satisfy simple commutation relations:

$$[a, a] = [a^\dagger, a^\dagger] = 0, \quad \text{and} \quad [a, a^\dagger] = 1\,. \qquad \text{(II.5.30)}$$

Furthermore we see that for the particle in a harmonic potential we can write the Hamiltonian as:

$$H = \omega(a^\dagger a + \frac{1}{2})\,. \qquad \text{(II.5.31)}$$

that they remain valid over time.

These operators raise or lower the energy of energy eigenstate with one step. This follows from the commutation relations:

$$[H, a] = -\omega a\,, \qquad \text{(II.5.32)}$$

$$[H, a^\dagger] = +\omega a^\dagger\,. \qquad \text{(II.5.33)}$$

Let us define the eigenstates $|n\rangle$ of the Hamiltonian as

$$H\,|n\rangle = E_n\,|n\rangle\,, \qquad \text{(II.5.34)}$$

then with (II.5.32), we obtain that applying $a^\dagger$ to a state $|n\rangle$, creates the state $|n+1\rangle$, because

$$H\,\{a^\dagger|n\rangle\} = (E_n + \omega)\{a^\dagger|n\rangle\},$$

and similarly for $|a\rangle$ with a minus sign on the right-hand side of the equations. Now we can see what we have gained with these manipulations. First we better assume that there is a lowest energy state $|0\rangle$ and as the energy cannot be lower we have to assume that the lowering operator gives zero when acting on this state:

$$a|0\rangle = 0\,, \qquad \text{(II.5.35)}$$

and thus:

$$H\,|0\rangle = \frac{1}{2}\omega\,|0\rangle\,.$$

From this one can show other quantessential properties:

$$E_n = (n + \frac{1}{2})\omega\,,$$

and,

$$|n\rangle = \frac{(a^\dagger)^n}{\sqrt{n!}}\,|0\rangle\,.$$

The results are summarized in Figure II.5.14. There are a few points worth mentioning. Firstly, the spectrum is equally spaced, and we have degenerate left and right movers. So it easy to construct raising and lowering operators. It is worth mentioning here already that later on in this chapter we will see an application of the oscillator algebra in field theory, where the operators $a^\dagger$ and a do not move

you through the spectrum of states of a single particle, but rather they act as creation and annihilation operators of particles in a given state, acting on a multi-particle Hilbert space. The second point is that the ground state has a non-vanishing 'zero point energy' equal $\frac{1}{2}\omega$, which basically follows from the uncertainty relations which do not allow the quantum particle be at rest at the bottom of the potential. The momentum (energy) cannot be zero. And indeed if you think of a table which is made up of zillions (or better 10^{25} or so) of oscillating particles you may wonder about the energy that appears to be just sitting there. In the cellar as it were, an incredible amount of vacuum, energy. What if....? May be we should just be cavalier about it and put in the same category as our friend the 'filled Dirac sea', where the physics basically only starts once you are on top, at the surface.

Constructing the wavefunctions. The explicit expressions for the wavefunctions $\psi_n(x) = \langle x|n\rangle$ are most easily obtained recursively starting from the ground state. The ground state wavefunction can be constructed by solving the equation (II.5.35) as follows:

$$\begin{aligned} & a|0\rangle = 0 \\ \Rightarrow \qquad & (\omega X + \frac{d}{dx})\psi_0(x) = 0\,. \end{aligned}$$

This is a differential equation with the (normalized) gaussian solution:

$$\psi_0(x) = (\frac{\omega}{\pi})^{1/4} e^{-\frac{\omega}{2}x^2}\,.$$

The higher states are obtained by repeatedly applying the raising operator $a^\dagger = \sqrt{\frac{1}{2\omega}}(\omega X - d/dx)$ on this ground state. So one just has to differentiate the ground state which is relatively easy to do. The resulting wavefunctions $\psi_n(x) = \langle x|n\rangle$ for the lowest n values were already displayed in Figure II.5.4 on page 383.

Coherent states

Let us return to the question of constructing quantum states that do look like a classical particle . These correspond to a *wave packet*, where we start combining waves in such a way that they have a reasonable width both in momentum and position space. We look for states that have a minimal spread about the average values of the variables, thereby making the uncertainty around a corresponding point in classical phase space in all directions as small as possible. Such states were already considered by Schrödinger and are nowadays called *coherent states*. They represent a wide class of states that just like the oscillator system have found many applications. These vary from quantum mechanics, optics, quantum chemistry, atomic physics, statistical physics, nuclear physics, particle physics, quantum information theory, group theory, and cosmology, to mention a few.

Let us now apply this idea to the states of a particle in the harmonic oscillator potential. We introduced the classical version of the harmonic oscillator already in the first chapter of Volume I on page 14. The periodic motion in configuration space that corresponds to a circular motion in phase space is characteristic. We now want to construct quantum states that show similar behavior. These cannot be the stationary energy eigenstates we have just been constructing in this subsection.

Minimal uncertainty states. From the commutator,

$$[\tilde{X}, \tilde{P}] = i\hbar\,,$$

directly follows the standard form of the uncertainty relation:

$$\Delta(\tilde{X})\,\Delta(\tilde{P}) \geq \frac{\hbar}{2}\,. \qquad \text{(II.5.36)}$$

What we would like to find is a state where we have that

$$\begin{aligned} & \Delta(\tilde{X}) = \Delta(\tilde{P}) = \Delta\,, \\ & \Delta^2 = \frac{\hbar}{2}\,. \end{aligned}$$

The states that achieve this are eigenstates of the lowering operator a, so we have:

$$a\,|\lambda\rangle = \lambda\,|\lambda\rangle\,, \tag{II.5.37}$$

these eigenstates have the basic property that $\lambda = \langle\lambda|a|\lambda\rangle$, but also that:

$$\langle E\rangle = \langle\lambda|E|\lambda\rangle = \frac{\omega}{2}\langle\lambda|(a^\dagger a + \frac{1}{2})|\lambda\rangle = \frac{1}{2}\omega(|\lambda|^2+1)\,.$$

Let us pause here for an instant. The question here is not to construct a ground state or an eigenstate of a given Hamiltonian, it rather is to construct eigenstates of the annihilation operator, with some eigenvalue λ. This problem is analogous to the construction of the translation operator that we discussed in equation (II.6.6). If we have an eigenfunction of position, where the expectation value of x is given by zero, we may apply the translation operator $T(a)$ to it, and shift the argument of the wavefunction so that $\psi(x) \to \psi(x+a)$. Then the vacuum expectation value will shift to $\langle x\rangle = -a$. And indeed the procedure is closely related, the desired state can be made out of the vacuum by a 'translation' operator built from the conjugate variable, in this case not the translation generated by the momentum P, but by $a^\dagger$:

$$|\lambda\rangle = e^{\lambda a^\dagger}|0\rangle\,. \tag{II.5.38}$$

So, if we write $a = (\tilde{X} + i\tilde{P})$, then such states have the property that:

$$(\tilde{X} + i\tilde{P})|\lambda\rangle = (\langle\,\tilde{X}\,\rangle + i\langle\,\tilde{P}\,\rangle)|\lambda\rangle = \lambda\,|\lambda\rangle\,. \tag{II.5.39}$$

Because for an eigenstate, the expectation value of the operator is equal to the eigenvalue. Bringing terms to the other side we obtain that:

$$|\tilde{X} - \langle\,\tilde{X}\,\rangle| = |\tilde{P} - \langle\,\tilde{P}\,\rangle|\,,$$

which establishes that the variances are equal: $\Delta(\tilde{X}) = \Delta(\tilde{P}) = \Delta$. From the equation (II.5.39)

$$\begin{aligned}\langle a^\dagger a\rangle = \lambda^2 &= \langle(\tilde{X} - i\tilde{P})(\tilde{X} + i\tilde{P})\rangle = \\ &= \langle(\tilde{X}^2 + \tilde{P}^2 + i[\tilde{X},\tilde{P}]\rangle = \\ &= \langle\tilde{X}^2\rangle + \langle\tilde{P}^2\rangle - \hbar\,.\end{aligned}$$

Taking the absolute square of equation (II.5.39), which contains the expectation values. This gives the result:

$$\lambda^2 = \langle\tilde{X}\rangle^2 + \langle\tilde{P}\rangle^2\,.$$

Combining the two previous results we obtain the equation for the sum of the variances:

$$\Delta(\tilde{X})^2 + \Delta(\tilde{P})^2 = 2\Delta^2 = \hbar\,,$$

giving $\Delta^2 = \hbar/2$ which is the minimum value allowed.

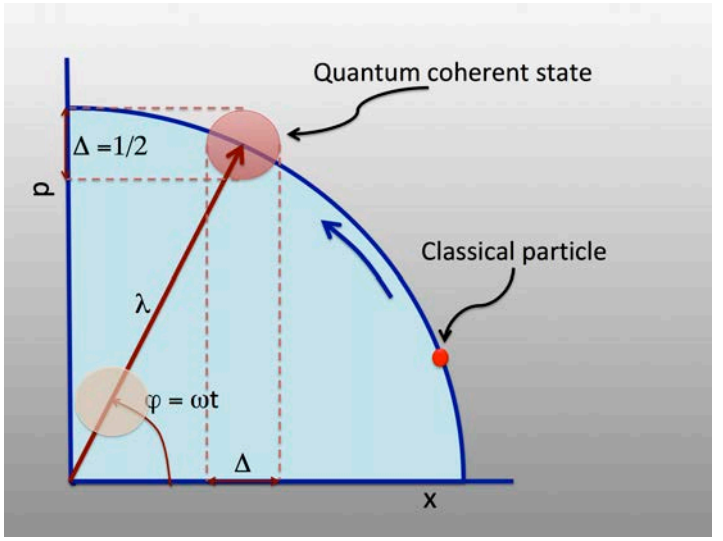


Figure II.5.15: *A fuzzy particle.* Phase space picture of the coherent state wavepacket with its fixed uncertainties for large values of λ. The coherent state x and p expectation values follow the classical trajectories but they carry a disk of uncertainties with diameter $\hbar/2$ along.

A fuzzy particle. What have we learned? Firstly that it is indeed possible to construct wave packets or coherent

states in which the uncertainties in position ϕ and momentum n match. In fact we found a continuum of different states $|\lambda\rangle$ that satisfy those conditions, and these states are labeled by the real parameter λ. Secondly we saw that average momentum is of order lambda, while the width of the momentum distribution in such a state is fixed and equal $\frac{1}{2}\hbar$. This means that if we increase λ the probability cloud of the particle becomes relatively narrow. The resulting overall picture is displayed in Figure II.5.15. The radial direction is the 'a' or therefore λ axis with real component x and imaginary component p. The time dependence of $a(t)$ is $a(t) = a\exp(i\omega t)$ as given in equation (II.5.29), so ωt is the angular variable in the figure. The resulting expectation values $\langle x(t)\rangle$ and $\langle p(t)\rangle$ describe the same trajectory in phase space as the classical particle would do. The classical periodic motion was depicted in Figure II.5.1 and the corresponding circular motion in phase space in Figure II.5.2 on page 380. We have emphasized that the uncertainties in position and momentum are fixed and independent of λ, which means that the approximation of the classical picture improves if we increase λ. This basically corresponds to the limit of high momentum or energy levels, where you would indeed expect classical behavior because the energies are large compared to the ground state level. However, note that as a function of time, the packet will broaden because the various momentum components move at different velocities. We depicted this type of broadening as a function of time in Figure II.5.3. ■■

The energy spectrum of coherent states. 
In this final paragraph of this section we show what the states $|\lambda\rangle$ look like if we decompose them in energy eigenstates. To do so we use a cute little trick. Note that the a operator, because of the commutation relation with $a^\dagger$, can be thought of as differentiation with respect to $a^\dagger$. This means that we can write:

$$a|n\rangle = a\Big(\frac{(a^\dagger)^n}{\sqrt{n!}}|0\rangle = n\Big(\frac{(a^\dagger)^{n-1}}{\sqrt{n!}}|0\rangle = \sqrt{n}\,|n-1\rangle\,.$$

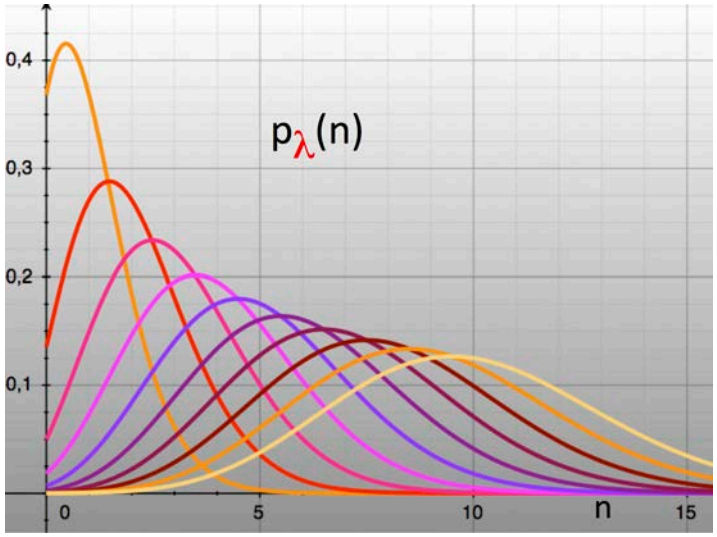


Figure II.5.16: *Coherent states.* The probability distributions $p_\lambda(n)$ given in equation (II.5.40) for finding energy $n \sim \lambda^2$ in a coherent state $|\lambda\rangle$ for $\lambda^2 = 1, \ldots, 10$.

The states $|\lambda\rangle$ can be obtained by finding a recursion relation for the coefficients α_n in (II.5.25) by imposing the defining equation (II.5.37):

$$a\Big(\sum_{n=0}^{\infty}\alpha_n|n\rangle\Big) = \sum \alpha_n\sqrt{n}|n-1\rangle = \lambda\Big(\sum_{n=0}^{\infty}\alpha_n|n\rangle\Big)\,.$$

Matching corresponding components we obtain the recursion relation:

$$\alpha_n = \frac{\lambda}{\sqrt{n}}\alpha_{n-1}\,.$$

This means that the states are given by:

$$|\lambda\rangle = N\sum_{n=0}\frac{\lambda^n}{\sqrt{n!}}|n\rangle\,,$$

with the normalization constant[11] $N = \exp(-|\lambda|^2/2)$. So what we have constructed here are coherent states parametrized by a parameter λ which have minimal and equal uncertainties for both conjugate phase space variables.

[11] Normalization of the state gives:
$\langle\lambda|\lambda\rangle = N^2\sum_{n=0}|\lambda|^{2n}/n! = N^2\exp(|\lambda|^2) = 1\,.$

These states have many momentum components; in fact we can calculate the energy distribution, for large λ it becomes:

$$p_\lambda(n) = |\langle n|\lambda\rangle|^2 = \left(\frac{|\lambda|^{2n}}{n!}\right) e^{-|\lambda|^2} . \qquad \text{(II.5.40)}$$

These are so-called Poisson distributions and we have plotted them in Figure II.5.16 for values $\lambda^2 = 1, \ldots, 10$. We easily calculate the average :

$$\langle n\rangle = \langle\lambda|n|\lambda\rangle = \langle a^\dagger a\rangle = \lambda^2 . \qquad \text{(II.5.41)}$$

whereas for the average of n^2 we obtain:

$$\langle n^2\rangle = \langle a^\dagger a a^\dagger a\rangle = \langle (a^\dagger)^2 a^2 + a^\dagger a\rangle\lambda^2 = \lambda^4 + \lambda^2 .$$

Combining the two we find for the variance:

$$(\Delta n)^2 = \langle n^2\rangle - \langle n\rangle^2 = \lambda^2 . \qquad \text{(II.5.42)}$$

We see that the the average of n is proportional to λ^2 while the width of the distribution goes like λ. This means that for increasing λ, the distribution relatively narrows . This is of course consistent with our calculation from the uncertainty relation (II.5.36) where we found the same variance. The resulting situation is summarized in Figure II.5.15. ■ ■ ■

Figure II.5.17: A 1962 conversation between Dirac (left) and Feynman (right) at a conference in Warsaw. (Source: Courtesy of Caltec Photo Archives.)

Fields: particle species

In this section on quantum fields we bring together a number of insights that we have touched upon in previous chapters. When saying field theory, we start by thinking about *free fields*, these are described for example by the Maxwell equations, the Klein–Gordon or the Dirac equation. All of them are relativistic wave equations and the question is what it means to *quantize* them.

Let us make some observations first.
(i) Fields are defined over all of space and they typically have an infinite number of degrees of freedom, and in that sense you can think of them as equivalent to an infinite number of particles.
(ii) You can think of the fields as being the generalized coordinates, meaning to say that the configuration space which for a single particle is just 'x'-space is now the space of field configurations.
(iii) In Chapter I.1 we have shown that for a field like the electromagnetic field we can define an energy and a momentum density and the logic of field quantization is to run the same program as before, and impose canonical quantization conditions for fields (as coordinates) and their associated momenta.

Thie procedure is quite involved and it took about thirty years before the first consistent field theory named Quantum electrodynamics (QED) was completed.

Field quantization. Without going through any calculations, which are generally quite messy and extensive, let me nevertheless give you some feeling for the results which are strikingly simple and beautiful.[12] And to transcend the swamp of words let me take the example of the simple scalar particle described by the Klein–Gordon field $\phi(x,t)$, which has to satisfy the relativistic equation

$$(\Box + m^2)\phi(x^\mu) = 0\,.$$

This has solutions which can be expanded as a sum of plane wave solutions with coefficients a and a^* look like:

$$\phi(x^\mu) = N\sum_{\mathbf{k}} \frac{1}{\sqrt{\omega_k}}[a_{\mathbf{k}} e^{ik^\mu} + a^*_{\mathbf{k}} e^{-ik^\mu}]\,,$$

with the definitions $x^\mu = (ct, \mathbf{x})$ and $k^\mu = (\omega_k, \mathbf{k})$ and moreover the K-G equation imposes $\omega_k = \sqrt{\mathbf{k}^2 + m^2}$. The coefficients have to be each other's complex conjugates to make the field real. In this case the momentum field would just be $\pi(x^\mu) = d\phi/dt$ which is indeed the time derivative of the 'coordinate' field.

Oscillators once more. In the present context the fields are the observables! So the quantum fields are operators, and as they are time-dependent, they are Heisenberg type operators. What that means is that in the above expression which is called mode expansion, the field ϕ on left-hand side becomes an operator, and on the right-hand side the operator property is carried by the coefficients. The modes are just the classical plain waves multiplied by operator coefficients $a_{\mathbf{k}}$ and their conjugates $a^\dagger_{\mathbf{k}}$. These act now like creation and annihilation operators. Performing the calculational gymnastics of imposing the commutation relations for the fields ϕ and π in the end boils down to commutation relations between the operator coefficients. The upshot is surprisingly simple:

$$[a_{\mathbf{k}}, a_{\mathbf{k}'}] = [a^\dagger_{\mathbf{k}}, a^\dagger_{\mathbf{k}'}] = 0\,, \qquad \text{(II.5.43)}$$

$$[a_{\mathbf{k}}, a^\dagger_{\mathbf{k}'}] = \delta^{(3)}_{\mathbf{k},\mathbf{k}'}\,. \qquad \text{(II.5.44)}$$

But now the air clears up! Compare this result with the commutation relations in (II.5.30). What have we got? We have obtained an infinite number of harmonic oscillators, each labeled by a momentum vector $\mathbf{k}$, and having a frequency $\omega_{\mathbf{k}}$. So, one (free) quantum field is equivalent to an infinity of oscillators and that rings an infinity of bells. The energy or Hamiltonian H of the field is *not* so surprising:

$$H = \sum_{\mathbf{k}} \omega_{\mathbf{k}}(N_{\mathbf{k}} + \frac{1}{2})\,,$$

with $N_{\mathbf{k}} \equiv a^\dagger_{\mathbf{k}}\, a_{\mathbf{k}}$, the so-called *number* operator. There is also a total momentum vector $\mathbf{P} = \{H, \mathbf{P}\}$ for the field:

$$\mathbf{P} = \sum_{\mathbf{k}} \mathbf{k}_\mu(N_{\mathbf{k}} + \frac{1}{2})\,.$$

The above equations naturally combine in an energy-momentum four vector P_μ for the field.

Multi-particle Hilbert space. And what does the Hilbert space for such a free field look like? Well, first we define a vacuum state $|0\rangle$ with the defining property that is annihilated by all $a_{\mathbf{k}}$ operators. Now we act with a creation operator on the vacuum:

$$a^\dagger_{\mathbf{k}}\,|0\rangle = |n_k\rangle \text{ with } n_k = 1\,.$$

This means that we have made a step in energy of $E = \hbar\omega = \sqrt{(mc^2)^2 + (\hbar\mathbf{k}c)^2}$, where I have put the constants back in. That energy corresponds exactly to the relativistic energy of a single particle of mass m with energy $E = \hbar\omega_k$ and momentum $\mathbf{p} = \hbar\mathbf{k}$. So, we are not raising the energy of a single particle. No, every time we work with an $a^\dagger$ operator we *create* an additional particle of the type described by the field in the corresponding momentum state.

[12] In this section we have set $\hbar = c = 1$ for convenience.

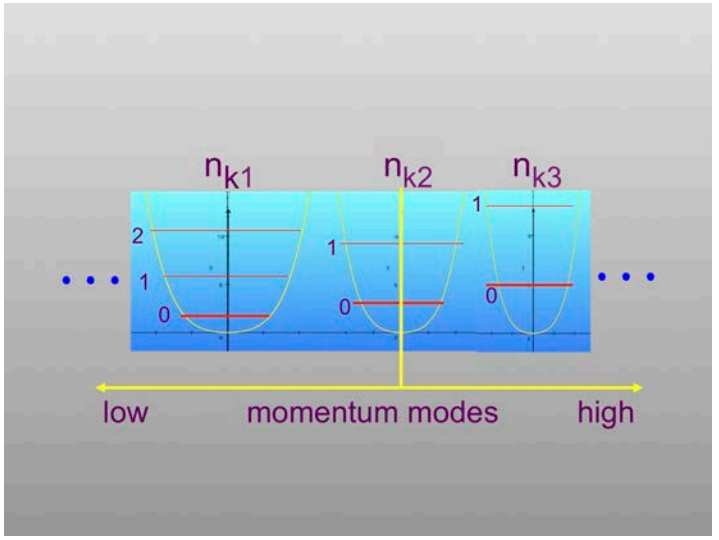


Figure II.5.18: *Quantum field modes.* A quantum state of a quantum field is labeled by the energy-momentum ($\vec{k}$) modes of the single particle, and the number of particles n_k that are in that mode.

And the annihilation operator does exactly the opposite. How charming and quantessential: the same algebra in another context creates another reality! The upshot is that we have a multi-particle Hilbert space, often called Fock space, with states,

$$|\{n_{\mathbf{k}}\}\rangle, \text{ with } N\,|\{n_{\mathbf{k}}\}\rangle = n_k\,|\{n_{\mathbf{k}}\}\rangle\,.$$

We have ended up with a clip and clear framework indeed.

The Klein–Gordon field is the simplest one to think of because it is just a field with one real component, but what about the other fields, like the Maxwell and Dirac field? Yes and no, their quantization is both similar but at the same time very different, also because their classical content is very different. In the Dirac case we have to understand what it means to have the Dirac sea and how to implement the anti-particles. Now the basic relations for the operators are *anti-commutation relations*,

$$\{b_{s,\mathbf{p}}\,,\, b_{s',\mathbf{p}'}\} = \{b^\dagger_{s,\mathbf{p}}\,,\, b^\dagger_{s',\mathbf{p}'}\} = 0 \tag{II.5.45}$$

$$\{b_{s,\mathbf{p}}\,,\, b^\dagger_{s',\mathbf{p}'}\} = \delta_{ss'}\delta^{(3)}_{\mathbf{p},\mathbf{p}'}\,, \tag{II.5.46}$$

and an identical set for the anti-particle creation and annihilation operators $d^\dagger_{s,\mathbf{p}}$ and $d_{s,\mathbf{p}}$. The index s denotes the spin state of the (anti-)particle. The anti-commutator is defined as the symmetric product, for example:

$$\{b^\dagger_{s,\mathbf{p}}\,,\, b^\dagger_{s',\mathbf{p}'}\} \equiv b^\dagger_{s,\mathbf{p}}\,,\, b^\dagger_{s',\mathbf{p}'} + b^\dagger_{s',\mathbf{p}'}\,,\, b^\dagger_{s,\mathbf{p}}\,.$$

This definition has a profound implication that becomes manifest if you look let the equation for a vanishing commutator work on the vacuum. It yields the result,

$$b^\dagger_{s,\mathbf{p}}\,,\, b^\dagger_{s',\mathbf{p}'}|0\rangle = -b^\dagger_{s',\mathbf{p}'}\,,\, b^\dagger_{s,\mathbf{p}}|0\rangle\,.$$

The two-particle states on the right and left have two particles in the same individual states but they are interchanged. We have interchanged two identical particles and that gives a crucial minus sign because of the anti-commutators. The relation with the Pauli principle becomes even more direct if you put $p' = p$ and $s' = s$, because then you get that that particular state equals minus itself, which means that that state is equal to zero! It says that such a state is just not there. It is not the ground state but a true no-state: a clearer statement of exclusion is hardly imaginable! With the Dirac equation everything fell into place: the spin appeared as necessary ingredient, along with the exclusion principle after the correct quantization. And then anti-matter as a bonus. How delightful! For the Maxwell field, it is the gauge invariance which has caused some profound headaches. But today all these difficulties have been overcome, and these type of (gauge) fields and their quantization form the basis of a consistent description of all particles carrying forces or interactions in the Standard Model.

Interactions. Of course if we discuss quantum field theory there is more than the quantization of free fields, it is a

multi-particle framework but the all-important interactions are left out. Isn't this about throwing out babies with the bathing water? No! This is a basic framework that is an absolutely vital starting point for any further going discussion.

Perturbative approaches. We have in Chapter II.1 already described some of the interactions that are present in the standard model. The basic interactions are characterized by certain interaction vertices, diagrams where different particles interact at a given space-time point. That point is where particles are annihilated and created in particular states that ensure that all the conservation laws like energy, momentum or charge, are respected. Each complete diagram then contributes to the overall probability amplitude for the process to take place.

This approach is called a *perturbative approach*, which is an iterative procedure to get ever better results, because in the calculations you include more and more complicated, higher-order diagrams. And as long as the coupling constant is small – and for QED for example the coupling strength is $\alpha = e^2/4\pi\hbar c \simeq 1/137$ – the higher order terms become tiny.

This way relatively low-order calculations already give incredibly accurate answers. And this scheme has led to the spectacular demonstrations of the power of quantum field theory, as for example in the calculation of the anomalous magnetic moments of the electron and the muon. The calculations are up to fourth order in α, and coincide with the best observed values up to 10 significant digits. This makes it the most accurately verified prediction in the history of physics!

Beyond perturbation theory. But in many situations it is necessary to go beyond perturbation theory. If either the particle density is large, or if the temperature gets very low, or the interactions become strong, one needs other approaches. And in the past century a lot of progress

The other currency

G: Hey Orange, I really like the stuff you told me about Dirac.
O: I am happy you liked it, Green. But you are right, he's a kind of a genius!
G: Yeah. That's what I thought, but more an anti-genius may be, chr chrr chrr!
O: He must have been very happy, with making discoveries of such profound importance for mankind.
G: Yeah. Hey Orange, I presume he must have become very, very rich.
O: You mean like Bill Gates or Warren Buffett.
G: or Prince or Picasso?
O: or Irving Stone or...
G: or Oprah!
O:Yes, you would think so Green. But no, I have to disappoint you.
G: But Orange, if you do such great works...
O:It didn't happen.
G: You mean that others have stolen his ideas?
O: No Green, it is not that. You have to understand Green, for scientific achievements like Einstein's of Dirac's or Heisenberg's there are no rights.
G: Are you telling me that they forgot to manage their copyrights or patents? These brilliant men didn't do their homework, is that it, chr chrr chrr.
O: Quiet down Green. Respect! Let me tell you this: a formula isn't like a novel, or a song, or baseball game, or a paperclip, or a diesel engine, or a talk show.
G: Are you saying that in the big scheme of things it is just marginal.

O: Yes indeed, Green, thank you. Now you understand what I mean.
G: Thank you Orange, I think I am going to have a peanut butter jelly sandwich! A Schrödinger-Dirac--Heisenberg sandwich! chr chrr chrrr.
O: Green! Listen, the scientist have another type of currency.
G: Like bitcoins?
O: Yes Green, but they call them *citations*.
G: What do those buy you?
O: Well, you know, Green, you know this game called monopoly? You can make a lot of money ...
G: I am getting really hungry. Thanks Orange. □

has been made in developing alternative non-perturbative ways of using field theory. We will discuss some important examples in the context of condensed matter physics in Chapter III.3.

Often situations where perturbation theory breaks down have to do with identifying some highly non-trivial ground state and start from there. For example it may be that a certain particle-type will condense in the ground state, so that it is no longer an eigenstate of the number operators $N_{s,\mathbf{k}}$. In fact one finds that some number density operator has a non-vanishing expectation value in the new ground state. The ground state of the super conductor is a canonical and beautiful example.

The phenomenon of superconductivity was discovered by Kamerlingh Onnes, but It took more than half a century to arrive at a really deep understanding of the underlying mechanism. Among other things the message to science seemed to be: 'Never give up!'

Let us briefly indicate what it means that the ground state of a physical system is characterized by some condensate. Think of the electrons in a conductor: they interact over relatively long distances via the lattice vibrations, which after quantization go under the name *phonons*. This phonon induced interaction between the electrons turns out to be attractive, and leads to a pairwise binding of the electrons of opposite spin and momentum. The electrons form so-called Cooper pairs. These pairs having spin equal zero, are of course bosons and therefore they can all condense in the same state. Indeed the ground state is a coherent state of Cooper pairs, which can be thought of as a linear combination of states with all possible different numbers of pairs in it. The system gains an enormous energy by dropping in this ground state, because the exclusion principle had pushed the individual electrons up to quite high energies. And starting from this ground state one has been able to prove all relevant properties of superconductors, using the successful BCS theory developed by the American physicists, John Bardeen, Neil Cooper and Robert Schrieffer, who received the Physics Nobel prize in 1964.

Ground states as coherent states. This situation is similar to the one we encountered in the previous section about the harmonic oscillator, looking at the phenomena of coherent states. In view of the almost uncomfortably close analogies between field theory and simple oscillators, it is imperative to ask about coherent states in field theory. What do they look like and what would the physics be like? Multi-particle coherence! What kind of bulk properties would that correspond to? And what low energy excitations would be there? Do we recognize them? What are interactions those 'trivial' agents could have engaged in, to give rise to such weird states? Here we enter a domain of what P.W. Anderson so beautifully characterized as 'more is different.' Many identical particles can, because of the interactions they have, give rise to highly non-trivial, highly diverse – but also highly non-recognizable – forms of collective behavior. Just like people, I am tempted to say. We have already encountered some of them, like quark confinement and the Higgs mechanism, but in the final part

of the book on structural hierarchies we will discuss many complex collective manifestations that emerge from the astonishing simplicity which we have exhibited here. The rich diversity of the condensed states of matter is the smashing consequence of having simple basic agents with simple basic interactions.

Particle spin and statistics

A quantessential principle with tremendous explanatory power is Wolfgang Pauli's exclusion principle*, decreeing that two or more Dirac-type particles (like electrons, neutrino's, or quarks) cannot occupy the same quantum state. Not all particles obey the principle, but if the particle does, it is called a* fermion*, and it also needs to have half-integral spin, just like the usual fermions described by a Dirac like equation. In this section we discuss a more direct and therefore more accessible approach to quantum statistical properties, based on the topology of the two-particle configuration space. The discourse is systematically built up, starting from the notions of indistinguishability and exclusion to describing particle interchange and the spin-statistics connection.*

Indistinguishability

> In quantum field theory, the loss of particle identity is inevitable

In quantum field theory the states correspond in general to many-particle states. These states are described by one field, or wavefunction, and this implies that individual particles are no longer distinguishable entities. A severe loss of identity in the quantum world. It is a world where only family names exist; first names are just not there.

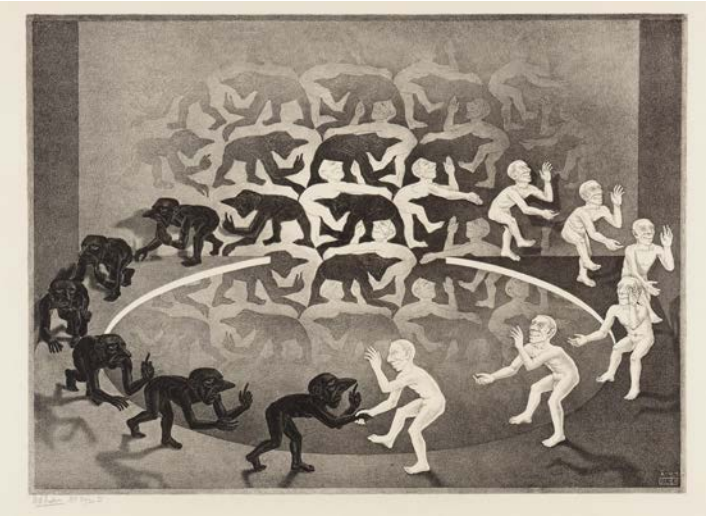


Figure II.5.19: *The Encounter.* This magical etching of Maurits Escher's was made in 1944. (© 2023 The M.C. Escher Company.)

The fact that multi-particle states are related to a single field implies an additional property, namely, that the corresponding particles loose their individuality. Individual particles of a given type, described by one type of field become indistinguishable. It may be that some state of an electron field describes two electrons, one electron in state A and one in state B, but you cannot say that particle 1 sits in A and particle 2 sits in B. They are like identical twins carrying a family name only but no first name. There is no 'John is at home' and 'Peter is at school', even though you *can* say that one is at home and the other at school. There is no 'who is who' in electron land (what a relief!), just strict anonymity and for that matter perfect democracy. Particles have a family name only. It may remind you of extremely strict school outfit rules: identical uniforms, identical shoes, and identical haircuts, in an attempt to wash away individual differences. Not my cup of tea. Anyway, this severe quantum loss of identity affects the counting of the available number of 'different states', and therefore the statistics properties of ensembles of such quantum particles. The statistical properties of the particles in turn are quantessential for understanding their collective be-

havior.

We will later return to the basic reason for that loss of individuality, being that all multi-particle/antiparticle states of a given species correspond to states of a single field describing that species, say the electron field or the photon field.

Exclusion

We have seen that quantization basically implies the study of wavefunctions of the classical configuration space. So we want to just focus on the special case that is of particular interest. Imagine that we have two particles that are 'identical', meaning that they are indistinguishable. These two-particles states are described by a single wavefunction defined on the two-particle configuration space, depending on the two position coordinates x_1 and x_2. But the indistinguishability of the particles implies that certain configurations which look different at first have to be identified. If somebody asks us to count the number of different (distinguishable) states, then we have to identify all configurations where the positions of identical particles are interchanged. Again, it's like a class where we have an identical twin, and we ask on how many different class configurations there are. Assuming that the twins are indeed indistinguishable by all means, we would have to count the state where twin A is in the front row and twin B in the back row and the configuration where they have switched places, as one and the same configuration. You see that the condition of indistinguishability affects the way we count the number of possible states, and therefore what the statistical weights are that we have to assign for certain configurations to occur.

There is however another important distinction we want to make right from the start. We may want to implement an exclusion rule saying that twins are not allowed to sit on the same chair. They may like each other but their sympathy is limited and sitting on the same chair is just out of the question. A rare occasion where the teacher and the twins seem to fully agree! Back to identical and indistinguishable particles, imagine the first particle has coordinate x_1 and the second x_2. The quantum state is then described by a two-particle wavefunction $\psi(x_1, x_2)$ depending on both coordinates. The question is now what we can say about the wavefunction if the two particles get interchanged, i.e. $\psi(x_1, x_2) \to \psi(x_2, x_1)$. Yes, their configuration is identical in that there is no experiment that can distinguish the two situations from each other – the usual nightmare for all twins. But does that imply that the wavefunctions have to be strictly equal? That's the question.

Unobservable phases? Taking into account all lessons we have been exposed to so far, we can say that the two wavefunctions can only differ by a subtle attribute that is not observable, namely the overall phase. It is subtle and seems completely innocuous but as we will see it is of crucial importance. This sounds indeed paradoxical, a supposedly unobservable phase that manifests itself. Let us first give the argument the naive and sloppy way, and say that the wavefunctions differ by a phase factor:

$$\psi(x_2, x_1) = e^{i\alpha}\psi(x_1, x_2)\,.$$

We expect that if we interchange them once more we will get back to the original state, from which it follows that we have to demand that:

$$e^{2i\alpha} = 1\,,$$

and this constraint has two solutions (modulo 2π) $\alpha = 0$ and $\alpha = \pi$. This in turn implies that there are two different solutions for the wavefunction under interchange of two identical particles:

$$\psi(x_2, x_1) = \pm\psi(x_1, x_2)\,,$$

implying that the wavefunctions are either symmetric or antisymmetric under the interchange. And indeed the particles that obey the symmetric rule are called *bosons*, the

antisymmetric guys are called *fermions*. We see that the antisymmetric solution implies that the particles cannot sit in the same spot, because if so that wavefunction would have to satisfy $\psi(x,x) = -\psi(x,x)$ implying that $\psi(x,x) = 0$! This 'unobservable' phase has huge quite observable consequences! This is so because the origin of *this* type of phase is topological.

Because of the indistinguishability requirement, the Hilbert space of two-particle states breaks up in two disconnected pieces being the even and odd functions. The phase is not the overall phase but the phase acquired under the interchange operation, and indeed the interchange should not change the observable probability distribution, which it doesn't.

Apparently fermionic particles obey an *exclusion principle* and such particles behave physically totally different from their bosonic counterparts, who are not subject to this exclusion principle and may like to hang out in the same spot. Indeed, they do like to sit on top of each other if it gets really cold!

Figure II.5.20: *Shinkichi Tajiri: Meandering paths (1997)* 'Meandering paths, unavoidably returning to an empty shell.' Looking at this work from a quantum perspective it depicts the entangled world-lines of particle pairs, first created and later annihilated. Indeed, the net effect is a transformation of the vacuum state. (Source: info@tajiri.nl.)

The topology of particle exchange

Two-particle configuration space. It will turn out that the possibility of non-trivial quantum statistics is directly linked to the connectivity properties of the configuration space of two identical particles and the topology of particle exchange. It is therefore worth considering in more detail what this 'two-particle configuration space' really looks like.

We start by taking two coordinates x_1 and x_2 which take values in some ordinary space $\mathcal{M} \sim \mathbb{R}^3$ for example. Instead of choosing x_1 and x_2 we may also choose as coordinates the 'center of mass' coordinate $X = (x_1 + x_2)/2$ and the 'relative coordinate' $x = (x_1 - x_2)/2$. During interchange we may keep X fixed (the origin, say), for example by moving the two particles around the center of mass that is located exactly half way between them. The interchange $x_1 \leftrightarrow x_2$ corresponds to a move from $x \leftrightarrow -x$ while keeping X fixed. So, we are left with studying the 'x' space. This space is again a copy of $\mathcal{M}$, but not quite, because in this space points, that are mirror images through the origin of each other, meaning the points x and $-x$ have to be identi-

fied if the particles are indistinguishable. Furthermore, the physical interchange in ordinary space corresponds to a closed loop in this reduced x space.

Three or more dimensions. We can take care of this doubling by cutting the space in half,[13] so we take away the bottom half of the space, say, all points with z negative $(z < 0)$, as we have indicated in Figure II.5.21. This solves the problem almost but not quite, because the space we are left with has acquired a bottom where strange things still happen. Indeed, in the bottom $z = 0$ plane we still have to identify the mirror points. But now this is at least something we can do 'by hand'.

Connectivity. The connectivity of the space is determined by studying the classes of possible loops in the space. Let us first discuss that and then return to the question of interchanges. In Figure II.5.21 I have drawn two paths. The first one is the green loop denoted σ, which is a loop that can smoothly be contracted to the red base point, hence it is the 'trivial' loop. This trivial loop means that there is basically no exchange and therefore the phase of the two particle state cannot change, so we conclude that $\sigma = +1$. The second red curve is again a closed loop because the beginning and endpoint are the same point, but now we can not contract the loop. The smooth deformations can only involve motions of the pair of red points into other mirror pairs in the bottom plane, if you were to lift them out of the plane they would no longer be the same point, and you would cut the loop – not so much a smooth deformation rather a killer move. And you cannot bring them together through the origin, because that point is taken out. So, the red loop is truly non-contractable and clearly belongs to a different topological class. We conclude that the reduced space clearly has some 'nontrivial' topology. The question is to find out what values the phase τ could take.

[13] 'Cutting the space in half' is not a typical act that experimentalist can perform. The point is that to make the topological argument we can do this in our head to simplify our analysis without loss of generality.

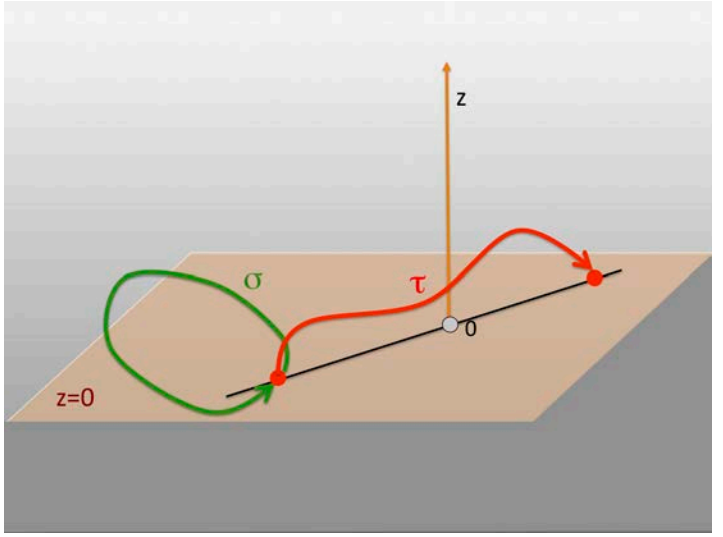


Figure II.5.21: *Topology of two-particle configuration space.* The two-particle configuration space, is $\mathbb{R}^3$ but with the bottom half and the origin removed. And on the $z = 0$ plane a point and its mirror image through the origin are identified. So there are two inequivalent types of closed paths possible. The green loop, which is contractable to a point, belongs to the trivial class; $\sigma = 1$. The red path, which is also closed but not contractable, belongs to the other, non-trivial class.

Interchanges As we said already an interchange $x_1 \leftrightarrow x_2$ corresponds to a move from $x \leftrightarrow -x$. Furthermore the path connecting the two points in x-space is not allowed to pass through the origin, because then they would meet at the same point and we would like to allow for an exclusion principle. An admissible move is depicted in the top graph of Figure II.5.22. In the reduced $x-$space this interchange is schematically depicted in the lower graph of the figure. We do allow the wavefunction to acquire some *constant* phase factor τ and that factor cannot change under a continuous deformation of the path from x to $-x$ through x-space. This means that the admissible phases τ label the different topological classes of closed paths that are possible in $x-$space. We have discussed these classes before, on page 83 of Chapter I.2, and learned that these are called *homotopy classes*.

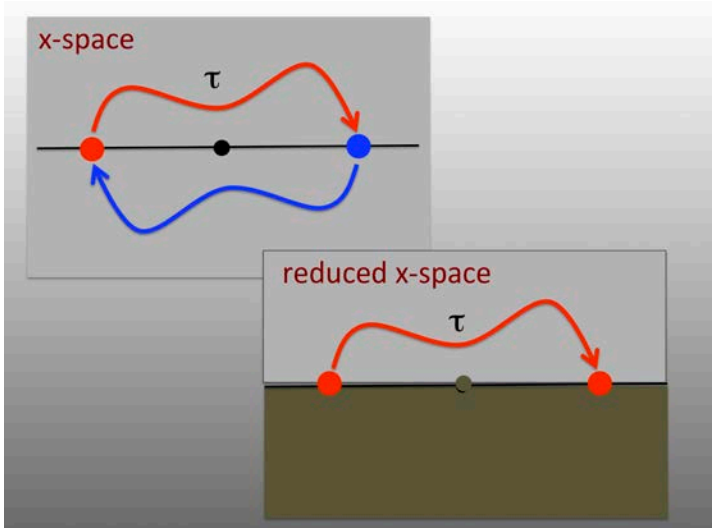


Figure II.5.22: *Interchange.* Particle interchange denoted by τ, in the space of the relative coordinate $x = (x_1 - x_2)/2$ amounts to moving from some point representing the pair, from x to -x along some path. In this particular case we have in fact that $x = x_1$ (red curve) $= -x_2$ (blue curve). The system has then moved to an indistinguishable two-particle state which means that the wavefunction can at most acquire a phase and we write $\tau\psi(x) = e^{i\gamma}\psi(x)$.

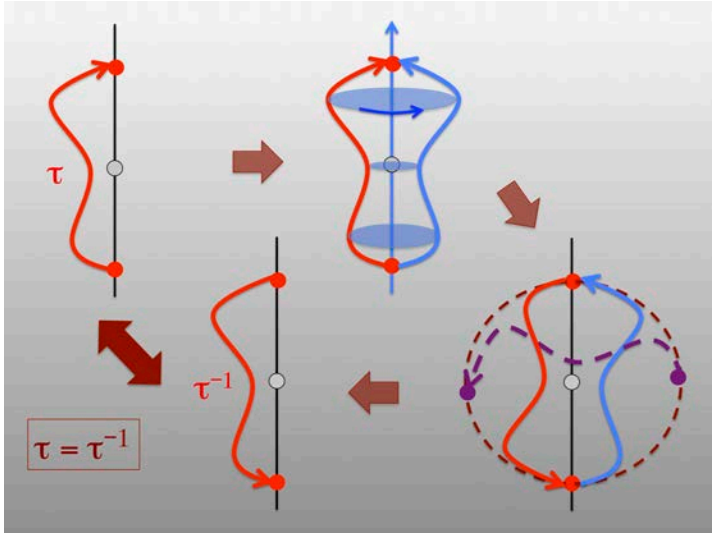


Figure II.5.23: *Topological equivalence.* The phase factor τ is the same for all interchanges along (closed) paths that can be smoothly deformed into each other. So τ labels a class of paths. In this figure we show that the class of τ and τ^{-1} are actually the same by a sequence of smooth deformations (rotations). Note however that the first move from red to blue is only possible if the dimension of the space is $D \geq 3$. In that case $\tau^2 = 1$ or $\tau = \pm 1$.

Let us now turn to Figure II.5.23 where we establish a relation between the interchange process τ and the reverse process represented by τ^{-1}. The top-left diagram is again τ and the bottom-left diagram represents by definition τ^{-1}.

Now we can do two subsequent smooth deformations of the path: in the top-right diagram we go from red to blue by just rotating around the blue axis, and in bottom-right diagram we go from blue to red again by rotating along the dark red trajectory indicated. Note that this deformation only involves mirror points (as is evident from the intermediate dark red dashed loop), so the loop remains closed and the origin is circumvented as required.

What we now learn from comparing the red path in the bottom-right diagram and the path corresponding to τ^{-1} is that these two paths can be smoothly deformed into each other, and therefore belong to the same class. The conclusion is that we have shown the surprising fact that $\tau = \tau^{-1}$, in other words that $\tau^2 = 1$, which implies that τ can only take the values $\tau = \pm 1$.

And therefore we confirmed that the quantum theory allows for only two fundamental types of particles: bosons with wavefunctions that are symmetric under particle interchange and fermions with wavefunctions that are antisymmetric.

But we also have the added restriction that the fermionic $\tau = -1$ solution requires the exclusion principle, corresponding to removing the origin of $x-$ space.

Finally, let us make a crucial observation that has been

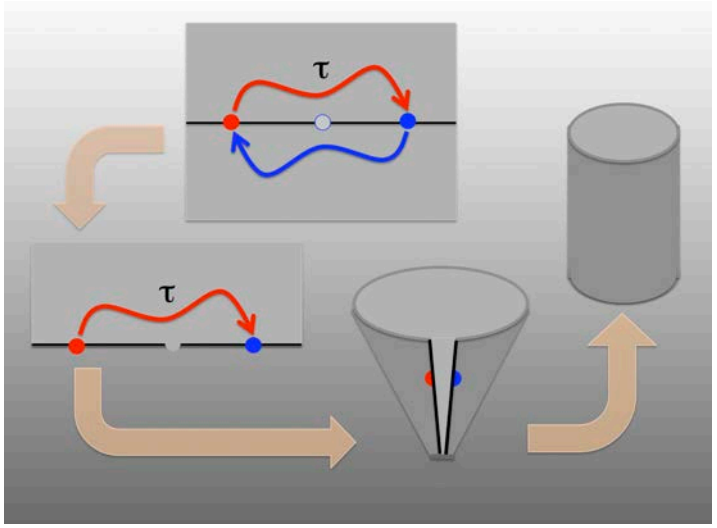


Figure II.5.24: *The two-dimensional case.* We start with the plane in which we do the interchange, the origin is excluded and we have to identify x and $-x$. This lets us remove the lower half of the plane. Then on the bottom boundary of the remaining top half space we still have to identify mirror points through the origin. This means the space becomes topologically a cone but without a tip. And that is topologically equivalent to a cylinder.

for a long time overlooked. The first 'red to blue' deformation can only be performed if the dimensionality of space is at least three; it requires $D \geq 3$! The question that remains is: what is so special about the two-dimensional case?

The two-dimensional surprise! In two dimensions the relative 'x' space is a plane with the origin taken out, and with opposite points identified. The paths of the particles are each other's mirror image just as we see in the top picture of Figure II.5.24. So again we go one step further and cut away the lower half-plane. Then we don't have to make any additional identifications except for points on the boundary. It is easy to visualize what that means. You can fold the half lines making up the boundary, together, literally by identifying the mirror points as indicated in the figure, and what you obtain is a cone! But: a cone without a tip. It is more like a *tipi* or an Indian tent with a hole in the top serving as a chimney to let the smoke out. Topologically speaking a cone without a tip is not a cone but a cylinder. And so, after all these topological moves we have shown that the space M_2 becomes an R^2 related with X, times a cylinder, $R \otimes S^1$, for x. The important conclusion is that interchanges in the original two-particle space $\mathcal{M}_2$, correspond to closed loops on this cylinder. And therefore the question of a topological characterization of 'identical' particle types is then reduced to the question of equivalence or homotopy classes of closed loops on a cylinder.

What we see is that the situation in two dimensions is special indeed, because we can imagine closed paths that wind around one time, two times, or n times around the cylinder and these are all inequivalent. So there is an infinity of classes which can be labeled by the set of (positive and negative) integers also referred to as winding numbers and denoted by $\mathbb{Z}$. And there is even a further property, you can compose loops, by joining end of the first loop (γ_1) to the beginning of the second (γ_2), then you get a combined loop ($\gamma_3 = \gamma_1 \cdot \gamma_2$). The corresponding classes of the loops will then add: $n_3 = n_1 + n_2$.

So in two dimensions it is in principle possible to have particles which satisfy $\tau^n = 1$ for any n, meaning that the phase factor of the two-particle state under interchange would be $\tau = \exp 2\pi i/n$. And that is why Frank Wilczek coined the generic name *anyons* for such particles because they evidently can have *any* phase.

And indeed, this observation would have the bold implication that in two dimensions the statistics factor could be any rational fraction of 2π, $\alpha = 2\pi/N$. By the ribbon argument which we explain in the next subsection, this would also imply that the spin value should be $s = 1/N$. How exotic: a correspondence between fractional spin and statistics!

Life in lower dimension is not always less interesting ap-

Figure II.5.25: *Feynman in discussion at the Les Houches Summerschool in 1979.* Feynman urged students including myself (who took the picture) to try and think of a simpler explanation of the exclusion principle.

parently! That can't be right! As a matter of fact, it is true, and there are states of matter on interfaces or with planar geometries where such particles exist. For example as collective excitations in (quasi) two-dimensional media like the 'fractional quantum Hall phases,' that are exhibited by certain conductors at extremely low temperatures, as we will discuss in Chapter III.3.

A historical aside. The topological nature of the particle exchange statistics goes back to work of the Norwegian Physicists Jon Magne Leinaas and Jan Myrheim from 1977. They applied the very same argument we employed in Figure II.5.22 and discovered the exceptional situation in two-dimensions. In 1980 I published a paper where I constructed explicit soliton solutions that exhibited fractional spin as well as (non-)abelian statistics properties. It was in the eighties that the extensions of these ideas took off within my own group, also guided by important developments in condensed matter theory such as the work of Laughlin and Wilczek on the fractional quantum Hall effect, and string theory and topological field theories by Witten. This has lead to a quite rich research field, nowadays called *topological order* or *topological matter*, in which these exotic features are realized and I myself was deeply involved. This research field is expected to have important applications in scalable and controllable quantum information processing and storage. And that is a good reason to explore these topological arguments a little further. It is an attractive type of physics, because it involves global analysis, which appeals to conceptual imagination rather than calculus type of skills. It's fun when basic (or fancy) physics meets basic (or fancy) mathematics; it really looks like these two fields of science are 'convicted' to each other. A marriage forced by nature on the one hand and a *marriage de raison* as the French say on the other, that should be a happy one.

The spin-statistics connection

We have in previous sections mentioned the remarkable connection between the fact that particles having half-integer spin happen to be fermions while the integer spin particles are always bosons. This spin-statistics connection between interchange properties and spin was not at all obvious from the start, and it only became clear once Dirac wrote down his famous equation for the electron and its anti-particle the positron that both properties were a necessary consequence of the brilliant interpretation of that equation given by Dirac.

But now we understand the topological argument for the interchange factor from carefully looking at the two- (or multi-) particle configuration space as we did in the previous section, one wonders whether there is not a more direct argument for the connection of this factor to the spin. There is, as we will show next, and it again turns out to illuminate the possibility of fractional spin for those aforementioned anyonic excitations.

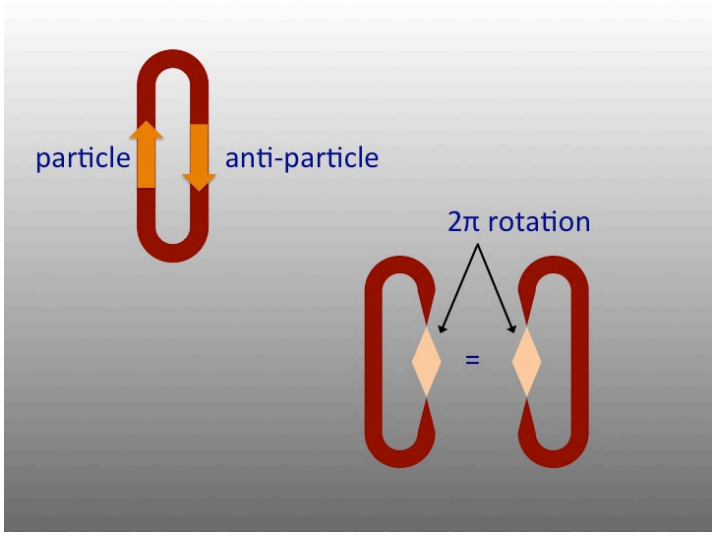


Figure II.5.26: *Ribbon diagrams.* The ribbon diagram of the creation and subsequent annihilation of a particle anti-particle pair, where the arrow indicates the direction of the charge current (left). The effect of rotation of a particle on the state is equivalent to the effect of a rotation of an antiparticle (right); the net effect is a change of of the vacuum state by a phase factor $R(2\pi)$.

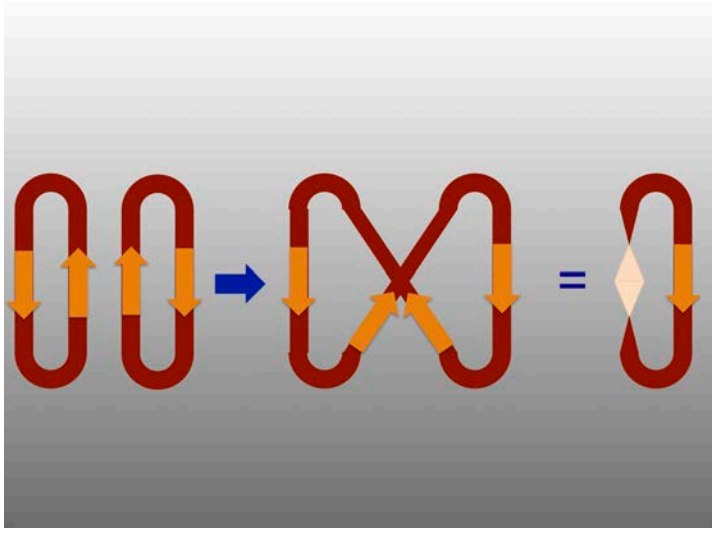


Figure II.5.27: *Spin – statistics connection.* Two pairs are created and annihilated, corresponding to a trivial effect on the vacuum state. The pictures on the right demonstrate the topological equivalence of the interchange of two identical particles with a rotation on one of them. This implies that $\tau\psi = R(2\pi)|\psi\rangle = \pm|\psi\rangle$, where the plus sign holds for *bosons* and the minus sign for *fermions*.

Ribbons. The trick is basically to realize that a particle with spin should be represented by a ribbon instead of a line. Let us imagine creating a particle anti-particle pair and subsequently annihilating it, then we get a diagram like in Figure II.5.26. We can of course also rotate the particle say over an angle of 2π before annihilating the pair, this corresponds to a full twist of the ribbon. What is demonstrated in the diagram on the right, is that we can move the twist smoothly from the particle line to the antiparticle line, which shows that their spin should equal. The rotation will change the phase of the two-particle wavefunction by an angle $\alpha = 2\pi s$ where s is the spin of the (anti-)particle.

To demonstrate the equivalence of a rotation by 2π to an interchange we go to the next Figure II.5.27. There we first create two pairs, then we cut the two identical particle ribbons and reconnect them to arrive at the diagram in the middle where the ribbons show that we interchanged the particles. In other words we have applied the interchange operator τ to the wavefunction describing the middle two particles. As indicated in the diagram on the right, the complete exchange diagram can be smoothly deformed into the diagram where one of the particles is rotated over 2π . This you can actually verify by taking a ribbon and literally repeat the described actions. What this says is the wavefunction of the state is acted on by the interchange operator τ shifting the phase of the state by an angle α, but this phase should be equal to $2\pi s$ according to the topological equivalence of the two diagrams.

So this simple argument nicely shows the topological nature of the statistics factor and of the spin-statistics connection. And who would have expected that you could give

#	A	B	C
1	1	2	
2	1		2
3		1	2
4	12		
5		12	
6			12
7	2	1	
8	2		1
9		2	1

Marbles distinguishable $\Rightarrow$ 9 states

#	A	B	C
1	x	x	
2	x		x
3		x	x
4	xx		
5		xx	
6			xx

Bosons indistinguishable $\Rightarrow$ 6 states

#	A	B	C
1	x	x	
2	x		x
3		x	x

Fermions indistinguishable exclusion $\Rightarrow$ 3 states

Table II.5.1: *State counting.* Counting states for 2 identical particles that can occupy one of three states. The tables list the possible 2 particle configurations for classical particles, bosons and fermions.

a 'ham handed' experimental 'proof' of the spin-statistics connection just using two identical belts!

Statistics: state counting

We return to the standard setting of more conventional quantum theory and illustrate how indistinguishability, exclusion, and interchange properties do affect the statistical properties of ensembles of particles. This becomes clear if one starts counting the available 'distinct' states.

Let us illustrate this state counting by considering a simple example of two identical particles labeled 1 and 2 that can be in either one of three states A, B and C. In the tables on the next page we have listed the distinct configurations for classical particles ('marbles') which are supposed to be distinguishable, for quantum particles that are indistinguishable but do not obey the exclusion principle (bosons), and for quantum particles that do obey the exclusion principle (fermions). Because the counting of available states is different allowing for 9, 6 and 3 states respectively, the probabilities are directly affected. For example assuming equal probabilities for each allowed state, one may ask a question like: 'What is the probability p that the two particles sit in the same state?' Clearly for the marbles the answer is $\mathrm{p} = 1/3$, for the bosons $\mathrm{p} = 1/2$ while for the fermions we have $\mathrm{p} = 0$.

For the case at hand we can define the two-particle state $\Psi_{\mathrm{ij}}(1,2) = \psi_{\mathrm{i}}(1)\psi_{\mathrm{j}}(2)$ as a product of the states of the individual particles where i and j could be A, B or C. We can thus think of Ψ_{ij} as a 3×3 matrix , for the classical states there indeed are $3 \times 3 = 9$ entries, for the bosons we have to require that the state would be symmetric $\Psi(1,2) = \Psi(2,1)$ corresponding to a symmetric matrix which indeed has 6 independent entries, while for fermions we have to require the state to be antisymmetric $\Psi(1,2) = -\Psi(2,1)$ corresponding to an antisymmetric matrix having only 3 independent entries because the diagonal ones have to be zero. Indeed, the state vector Ψ where the fermions would be in the same state would

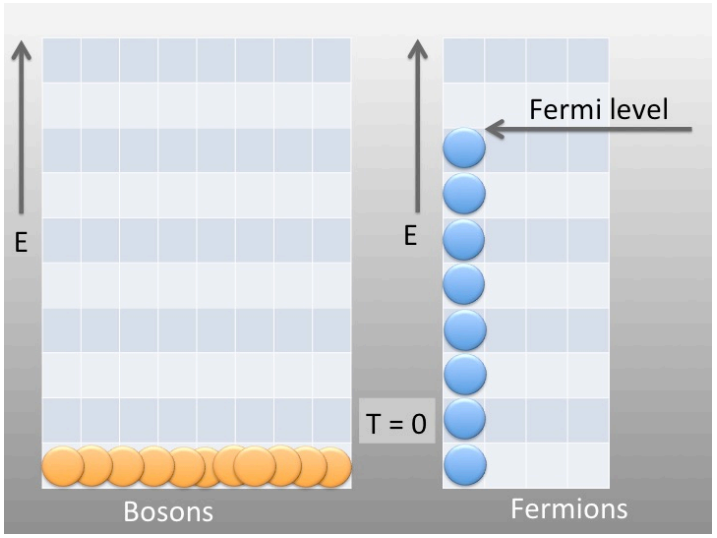


Figure II.5.28: *Bosons and fermions at* $T = 0$. The distributions for the two particle types at $T = 0$. The energy levels are along the vertical axis, and the occupation number is indicated by the number of balls.

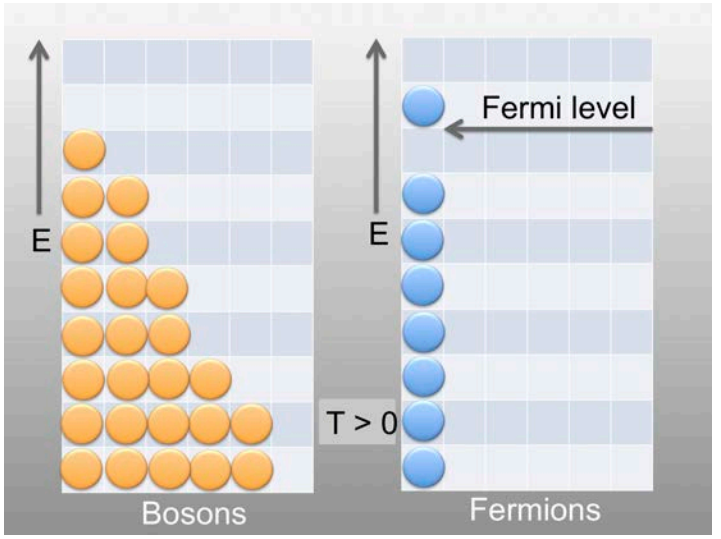


Figure II.5.29: *Bosons and fermions at* $T \geq 0$. Axes are the same as in previous figure.

mean $\Psi_{ii} = -\Psi_{ii}$ implying that it has to vanish, saying nothing less than that that there is no such state.

These basic statistical properties of particles have profound physical consequences if we study many particle systems and their collective behavior. For a system in thermal equilibrium with its environment, there will be a certain probability of a certain energy level to be occupied or not, which means that in a large system of many particles you get a distribution which tells you how many particles there will be on average at a certain energy level. Now dependent on the type of particle, these distributions are different, especially if one goes to low temperatures and low energies where the quantum behavior becomes manifest.

What do we roughly expect to happen? Let us start with taking the zero temperature case, this is shown in Figure II.5.28. Indeed for the bosons we expect that they all congregate or better condensate in the ground state. This is in contrast with the fermions where we expect that for N fermions, the lowest N states would be filled, while the higher states would be empty. The highest filled level is called the Fermi level, corresponding to the Fermi energy. Now if we heat the system up, particles may get excited to higher levels, and fall back again until a certain temperature dependent distribution over states is reached. So, in Figure II.5.29 we have indicated what that looks like. Clearly for the fermions where all lower levels are filled already the thermal excitations can only take place near the Fermi level. Fermionic excitations create in fact also a hole, near the Fermi level one necessarily creates particle-hole pairs.

The functional form of the three distributions can be determined exactly, and are are depicted in Figure II.5.30 for two different temperatures. They have the following functional form:

$$n_T(E) \sim \frac{1}{e^{(E-\mu)/kT} + m},$$

where for m=0 we have the classical Maxwell–Boltzmann distribution corresponding to the blue curves, while for $m =$

$+1$ we have the Bose–Einstein distribution corresponding to the red curves, and finally for $m = -1$ the Fermi–Dirac distribution corresponding to the dark red curves. You may think of these distributions as function of particle state energy, parametrized by the temperature and the chemical potential (Fermi energy) denoted by μ. Let us make some observations concerning these distributions.

i. Note that the axes in Figure II.5.30 are labeled orthogonally to those in Figures II.5.28 and II.5.29.

ii. Observe that for high enough energy all the distributions look the same for all temperatures, which is the statement that all particles approximately show the classical behavior. The quantum distinctions get washed away by the violent thermal fluctuations.

iii. Drastic differences however show up for low values of relevant energy scale $E-\mu$. Whereas the fermion occupation number necessarily is smaller than or equal one, the boson occupation number increases rapidly if the energy goes to zero. In fact, if we lower the temperature to absolute zero the fermion distribution function becomes a step function indicating that up to the Fermi-level, all states are occupied (here μ is the fermi-level, or the surface of the Dirac sea). For bosons we see that all particles will pile up in the same lowest energy state.

iv. There is actually a real phase transition where a so-called Bose-condensation takes place where all particles sit in the quantum same state. This is in fact an example of a special *macroscopic quantum state* that stands out because of its so-called quantum coherence. Such states exhibit truly spectacular properties, such as superfluidity, meaning that the system forms a quantum fluid with zero viscosity. In certain metals this can lead to the phenomenon of superconductivity, where the electric resistance vanishes at very low temperatures. We will return to these subjects in later chapters.

You may wonder how such peculiar rules like exclusions and indistinguishability can be implemented in a mathematically consistent way. It turns out that to do multi-particle (often called many body) quantum physics, you basically

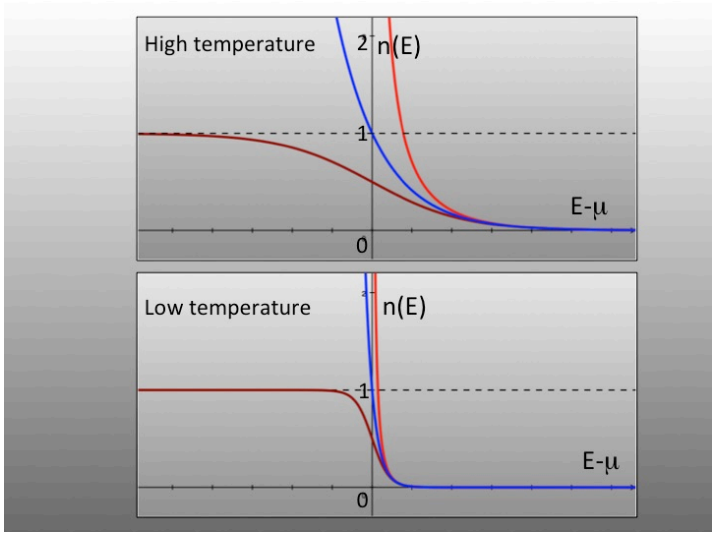


Figure II.5.30: *Particle distributions.* The distributions for three particle types, giving the occupation number $n(E)$ of a state at energy $E-\mu$ for two temperatures. The red curves are for bosons, the blue ones for 'marbles' and the dark red ones for fermions. This figure is rotated 90^0 clockwise with respect two the previous figures.

have to use the formalism of quantum fields. In this formalism we have operators that can create or annihilate (anti)particles in any admissible energy-momentum state. And one finds that the different types of statistics are direct consequence of the basic relations between these particle creation and annihilation operators. For bosons we that the creation and annihilation operators satisfy *commutation relations* meaning that

$$[a_k^\dagger, a_{k'}^\dagger] = 0\ ;\ [a_k, a_{k'}] = 0 \text{ and } [a_k, a_{k'}^\dagger] = \delta_{k\,k'}\,,$$

where the commutator of two operators A and B is defined as $[A,B] = AB - BA$. For fermions these are replaced by *anticommutators* where the anti-commutator is defined as $\{A,B\} = AB+BA$. If two creation operators anti-commute one has in particular that

$$\{c_k^\dagger, c_k^\dagger\} = 0\,,$$

meaning that putting two particles in the same state gives

zero, it just can't be done. This necessary choice of commutation or anti-commutation relations for the basic operators is forced upon you by the requirement of a physically consistent interpretation of the theory. That choice accounts for all characteristic differences between bosons and fermions in particular the appearance of completely symmetric or antisymmetric wavefunctions.

More for less: two-dimensional exotics

> It is like the telephone game in kindergarten. The children are sitting in a circle and you whisper the first kid a sentence in her ear, then she has to pass it on till it went all the way around. The last person speaks out loud what the sentence was he received. Then they compare the sentences, and share their unbelief that such distortions are possible. That is presumably how lies emerge. This metamorphosis, amounts to a non contractable loop in language space, a nontrivial linguistic holonomy.

The Aharonov–Bohm phase. We recall the discussion we had in Chapter II.3 on the Ahoronov–Bohm phase shift. If you carry a charge q along a loop γ, around localized flux then the loop integral of A along γ yields the magnetic flux through (any) two-dimensional surface hat is bounded by the loop. This implies that the loop operator W_γ basically measures the magnetic flux:

We considered a well-defined narrow magnetic flux tube piercing through the surface as in Figure II.5.31. If we adiabatically move a charge around the flux Φ, the state will change according to,

$$|q,\Phi\rangle \to W_\gamma(q,\Phi)|q,\Phi\rangle\,,$$

where the phase factor W equals

$$W_\gamma(q,\Phi) = e^{iq\Phi}\,.$$

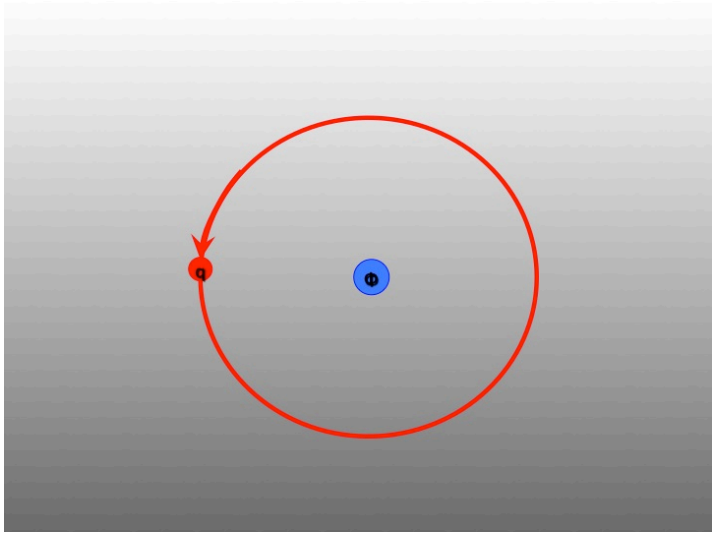


Figure II.5.31: *The Aharonov–Bohm phase factor.* If we carry a charge q along a loop γ around a localized magnetic flux Φ, then the state will acquire a phase factor $W_\gamma = \exp iq\Phi$.

An important property of this phase is that it is not only gauge invariant but also topologically invariant, meaning that you can deform the loop any way you want as long as you don't cross the flux.

Anyons as flux-charge composites. Let us return to our discussion about two-dimensional particles and their spin and statistics properties. Let us look once more at Figure II.3.33 but in a different way. I now think of the charge and flux as one composite object. The situation is like in Figure II.5.32, where we look from far away and do not worry about the (internal) structure of the pair. The interpretation of the figure is then that we rotate the composite over an angle of 2π, and we see that the state of this funny particle has changed by $W(q,\Phi)$. This means that our conclusion has to be that the composite must carry some spin s, which causes the non-trivial phase factor of the state under rotation by 2π. By definition for a particle carrying spin s, the corresponding factor is given by,

$$e^{2i\pi s} = e^{iq\Phi} \quad\Rightarrow\quad s = q\Phi\,.$$

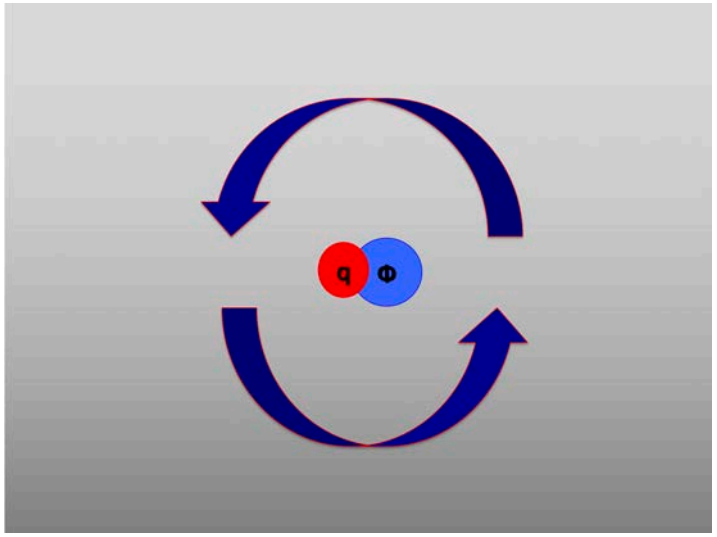


Figure II.5.32: *Flux-charge composite.* We think of the charge flux pair as a composite particle. Then the electromagnetic phase factor can be interpreted as due to a '(fractional) spin' s of the composite.

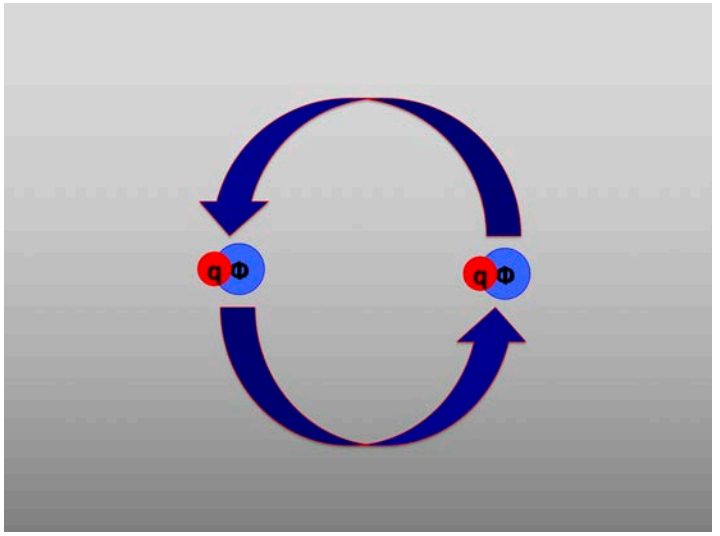


Figure II.5.33: *Interchange statistics of composites.* For the composite particle it follows that the spin- statistics connection holds.

For example, if in the superconducting layer, a single electron would bind with a minimal flux ($\phi_0 = \pi/e$) we would have $s = e\phi_0/2\pi = 1/2$, this would be a spin-half composite particle!

The spin-statistics connection for composites. We argued that the composites can have fractional spins depending on which fluxes and charges are allowed. But is it also true that they would exhibit the corresponding exchange properties? Can we establish a spin-statistics connection using the ribbon diagrams of Figure II.5.27 ? Let us start with the phase factor of two composites as in Figure II.5.33. The combined state after a full rotation would obtain a phase factor of twice $W(q, \Phi)$, because the charge q_1 would encircle the flux Φ_2 and at the same time q_2 the flux Φ_1, giving us $2q\phi$ as the fluxes and charges are equal. So we have to take the square root, as we only want to do the interchange, so we do get indeed the same result as the spin factor.[14] This way we have established the exotic spin and statistics properties that are possible in two dimensions.

A historical aside. These particles are called *anyons*, a name coined by Frank Wilczek, because they can acquire any phase upon rotation or interchange. These so-called quantum-Hall systems were discovered by the German physicist von Klaus von Klitzing, and the fractional version of it by Störmer and Tsui. The theory of this phenomenon involving the fractionally charged anyons with fractional spin along the lines we just pointed out was developed by the Americans Robert Laughlin who shared the Nobel prize with Störmer and Tsui in 1988, and Frank Wilczek who already had received a Nobel prize for the theory of the strong interactions.

There are now many proposals for phases of condensed matter that feature these local anyonic excitations. Such phases share a property called *topological order*. It was

[14] The possible extra minus sign from taking the square root cannot be resolved at this level of the analysis. Note however that it allows for implementing that the constituents would be a fermion to start with.

the Russian theoretical physicist Alexei Kitaev who pointed out that such anyons would be ideally suited to build quantum information devices with, because anyonic qubits are intrinsically fault tolerant. This highly desirable property derives from the topological nature of the quantum phases, which makes that these cannot be destroyed by local interactions and such error generating effects would be exponentially suppressed. One may manipulate the phases on multi-anyonic, multi-qubit states by just moving them around each other, or as it is called by 'braiding' them. Because of their topological nature computations with anyons would correspond to particular braids or knots of their world lines. And computation would boil down to some kind of quantum knitting! ■

Textbooks on particles and fields

- *Particles, Fields and Forces: A Conceptual Guide to Quantum Field Theory and the Standard Model*
  Wouter Schmitz
  Springer (2019)
- *The Quantum Theory of Fields*
  Steven Weinberg
  Cambridge University Press (2013)
- *An Introduction To Quantum Field Theory*
  Michael E. Peskin and Daniel V. Schroeder
  CRC Press (1995)
- *PCT, Spin & Statistics, and All That*
  Ray F. Streater and Arthur S. Wightman
  Princeton University Press (2000)

Chapter II.6

Symmetries and their breaking

> Symmetry, as wide or as narrow as you may define it, is one idea by which man through the ages has tried to comprehend and create order, beauty and perfection.
>
> Hermann Weyl

Symmetries play and have played a crucial role in the development of the modern physical sciences. It is a rich subject and its manifestations are quite diverse and display remarkable analytical and aesthetic aspects. Central to this topic are the mathematical notions of a Lie group *and a Lie algebra. In the quantum context these symmetries are implemented by certain sets of operators (observables) that act on the Hilbert space of the system. We have encountered them already as they arise naturally at many levels in the framework of quantum theory. The connections between formal mathematical and physical concepts are summarized in the table on page 447, and I recommend that you regularly consult the table while reading this chapter.*

In this chapter we have split the applications between the well-known 'ordinary'*,* rigid*, or* global symmetries *and the so-called* gauge *or* hidden *or* local symmetries*. The former are like the familiar translations or rotations, or isospin transformations, while the latter refer to the internal symmetries that are tied in with the fundamental interactions. Electrodynamics is a simple example of a gauge theory, and we have already discussed itsgauge symmetry already in Chapter I.1. Gauge symmetries are especially powerful because they are restrictive in the sense that they impose the way particles can interact in a consistent way. The dynamical equations underlying the Standard model are pretty much an expression of this principle of local gauge invariance. The mathematical concepts are those of differential geometry and the theory of fiber bundles, as we pointed out in the section on the 'Physics of geometry' of Chapter I.2*

After the discussion of symmetries themselves, we move on to talk about breaking the symmetries. Symmetry breaking *is another powerful concept that has found a rich variety of applications in fundamental physics on all scales, from say the cosmos all the way down to the phenomena of ferromagnetism in condensed matter or the Higgs mechanism in particle physics.*
Symmetry breaking encompasses a hierarchical perspective on the increasing diversity and complexity we observe in nature as a hierarchical pattern resulting from a sequence of symmetry breaking transitions. We will discuss examples of the breaking of global as well as local symmetries.
Symmetry and its breaking are deep and delightful subjects that teach us about the mathematical intricacies of fundamental interactions and their structural beauty.

Let me start this chapter by stepping back and revisiting some statements I have made along the winding road we have taken so far, and looking at them again from the point of view of symmetry. Symmetry pops up everywhere and that indicates that there are many entries into this quantessential subject. Whereas symmetry leads to unity, similarity, and degeneracy, breaking symmetries does the opposite, it is a mechanism explaining how symmetry can get lost. The mechanism is quite generic and it is therefore important to understand its systemic signatures.

Nature started from a highly degenerate situation at a very high temperature (energy) and then created (evolved) diversity by going through a series of symmetry breaking transitions that took place when the ambient energy or temperature lowered. In an expanding universe like ours the loss of symmetry is as natural as it is inescapable.

By changing a circle into an ellipse and then to an arbitrary closed curve, one goes from a symmetry of continuous rotations in the plane, to two mirror symmetries, to no symmetry at all. It is a sequence of ever more symmetries being broken. Note however that from an information point of view, the information content increases with decreasing (or the breaking of) symmetry. Indeed you move from a curvature along the closed curve that is constant, to a curvature that is a periodic to a random function, and the amount of data you need to describe them increases.

Too much symmetry is boring because it is extremely redundant and predictable, but the same holds for too much randomness because of an extreme lack of structure. Excitement and beauty apparently reside halfway in between, and that is maybe why nature has chosen a path of breaking more and more symmetries. At present we encounter remnants of lost symmetries like subtle and hidden memories. But that is what makes nature so interesting. Life as an 'avenue of broken symmetries' so to speak. It allows science to gain a deeper and more unified understanding of the hidden patterns underlying reality.

Symmetries of what?

> The symmetries that are important in physics, are not the symmetries of things but the symmetries of equations.
>
> *Steven Weinberg*

We think of a group of symmetries as a set of operations or transformations that leave something invariant. This can be an object like a triangle or a sphere, and we speak of the '*symmetries of objects*', and this is certainly its most familiar manifestation. We may also think of the *symmetries of spaces*, these are transformations on the space, meaning transformations of the coordinates in such a way that the properties of that space do not change. For example flat space $\mathbb{R}^3$ has a huge group of symmetries: we can translate it over an arbitrary distance in any direction, we can rotate it around any axis through any point over any angle, and we can scale it by any amount around any point. With an infinite flat space you wouldn't see the difference, it is invariant under all those transformations and combinations of them. And besides that it has also discrete mirror symmetries, a transformation called *parity*. It makes you wonder whether it is this incredible overkill of symmetry that makes flat space so boring.

Yet another, and in physics crucial, application is to study not so much the symmetries of things, but rather the *symmetries of equations*, which means again that we make a transformation on the dynamical variables that leave the (system of) equations invariant.

Realizations of symmetry in nature. People I trust have told me that the Inuits have 32 words for snow, and that presumably is because they know a lot more about it than I do. By living in the snow for centuries they have learned to differentiate and appreciate an immense diversity in something that I just call 'snow.' Something similar has happened with the notion of symmetry in physics and its mirror

images in mathematics.

With all these different approaches comes a correspondingly rich terminology referring to what we are precisely talking about. One speaks of *discrete* versus *continuous*, *finite* versus *infinite*, *space-time* versus *internal*, *local* versus *global*, *broken* versus *unbroken*, *approximate* versus *exact*, *normal* versus *super*, *classical* versus *quantum* symmetries. This summary suffices to justify a chapter on this topic, a chapter in which I will guide you through some of this extensive jargon in a way that emphasizes the basic concepts.

Groups, algebras and their representations. The framework for the following discussions on symmetry is summarized in the table on page 447, and it shows that in the class of continuous symmetries the mathematics is mostly that of Lie groups and algebras. These are quite abstract, mathematically precisely defined objects themselves, but the beauty is that it comes with an important part denoted as representation theory. Physicists perceive the notion of symmetry mostly through the particular representations that are manifest in nature. Let me recall the observables $\{X, Y, Z\}$, the Pauli matrices, and the fact that their commutation relations form the non-commutative Lie algebra denoted as $\mathfrak{su}(2)$.[1] It is called the 'defining' representation of this algebra because it is in the form of 2×2 hermitian matrices, working on a two-dimensional complex vector space – the state space of a single qubit. But exactly the same algebra, meaning an identical set of commutation relations, is obeyed by the angular momentum operators $\{L_x, L_y, L_z\}$. That is a different representation of the same algebra in terms of differential operators working on a space of functions – the Hilbert space, quite different from 2×2 matrices but satisfying the same algebra. If we furthermore restrict to states of a given angular momentum l, (think of the hydrogen atom) then these form a $(2l+1)$-dimensional vector space and the rotations are then generated by a specific set of three $(2l+1) \times (2l+1)$ hermitian matrices. And all these sets form inequivalent representations of the same algebra, labeled by the quantum number l. We will be somewhat cavalier about making distinctions between the abstract notions of an algebra or group and their representations. In physics we mostly work within the context of particular, often unitary, representations. You may think of representation theory as the physical contextualization of abstract group theory.

[1] To be precise, it is one-half times the Pauli matrices that satisfy the $\mathfrak{su}(2)$ algebra. Commutation relations are nonlinear so the scale is exactly fixed. This factor one-half turns out to be important.

Symmetries and conserved quantities

Heisenberg equations. I choose a route that starts with symmetries of a Hamiltonian (operator), leading from there to the notion of conserved quantities, and from there to frameworks for labeling the energy eigenstates of that Hamiltonian. Let me start from the basic Heisenberg equations which apply to quantum systems on all levels:

$$i\hbar \frac{dA}{dt} = [A, H] \,. \tag{II.6.1}$$

Remember that in this formulation the dynamical variables or observables are time dependent, and in that sense the Heisenberg approach is closer to the classical one, because it is formulated in terms of the observable quantities only.[2] This in contrast with the Schrödinger equation which describes the time evolution of quantum states, and those are not directly observable.

Symmetries and conservation laws. The equation says that the time evolution of the system is generated by the Hamiltonian H. In particular, an infinitesimal change in time, corresponding to acting with $i\hbar d/dt$ on the variable, is equal to taking the commutator of that variable with the

[2] Note the similarity between the Heisenberg equations and the Poisson equations discussed in the section on classical mechanics of Chapter I.1.

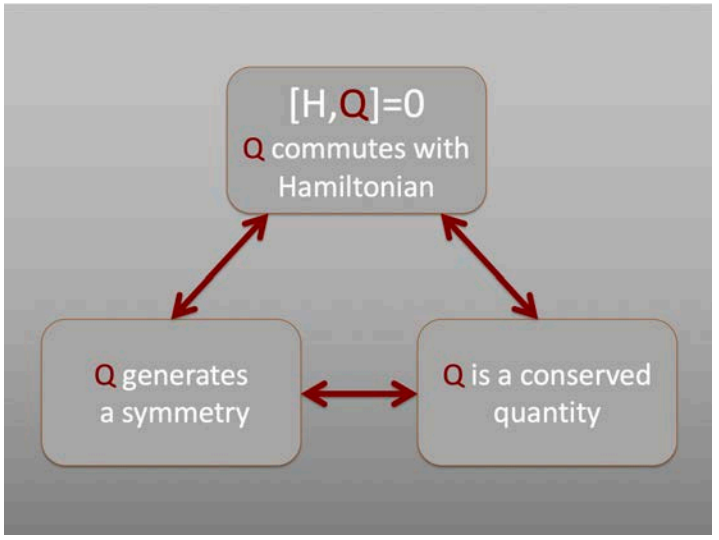


Figure II.6.1: *The quantessence of symmetry.* If an observable Q commutes with the Hamiltonian, then it is conserved in time, and generates a symmetry of the system.

Hamiltonian. Consider now an observable Q_i which commutes with the Hamiltonian or energy operator, so:

$$[Q_i, H] = 0 \quad \Rightarrow \quad \frac{dQ_i}{dt} = 0\,.$$

The equation teaches us that observables that have a vanishing commutator with the Hamiltonian do not change in time. They are constants of the motion and are conserved in time. It means that if the dynamics of the system follows the Heisenberg evolution equations, and we start with a state corresponding to a certain (eigen)value q_i for the observable Q_i, that the evolution will take place in a subspace of the Hilbert space labeled by that eigenvalue, and by some eigenvalue E of the Hamiltonian as well, because the energy operator H is (by definition) a conserved quantity. Everybody commutes with themselves after all.

This reasoning leads to an interesting picture: we have a system characterized by a set of basic variables (think of position and momentum) and a huge set of derived observables (like energy or angular momentum), and these observables form a closed operator algebra under commutation. In the *Math Excursion* on vectors and matrices on page 632 of Volume III we explain that these algebras of observables that close under commutation are in mathematics referred to as *Lie algebras*. We present an overview of the relation between mathematical and physical aspects of symmetry in the table on page 447.

Lie algebra of observables. What we say is that such a Lie algebra is a rather abstract thing, but it has representations in the form of matrices or differential operators. This we saw for example with the algebra of the canonical variables X and P, which reads:

$$[X, P] = i\hbar \quad \Rightarrow \quad X \to x \;\text{ and }\; P \to -i\hbar\frac{d}{dx}\,,$$

and therefore has a representation where X is represented by the ordinary number variable x (like it appears as argument of the wave function). Acting with X on a wavefunction $\psi(x)$ means multiplying that wavefunction with x. P is represented by the differential operator as indicated in the equation above. It is the infinitesimal displacement operator. This was worked out in the section on position and momentum operators on page 387.

Translation invariance and momentum conservation. Let us explore this a little further along the lines of energy conservation for the simple mechanical system that we discussed in the section on Newtonian mechanics in Chapter I.1. If we consider the energy of a particle then that usually consists of a kinetic part $P^2/2m$ and a potential part $U(X)$. Suppose that we make the additional assumption that the potential energy is constant and does not depend on X, then the canonical commutation relations above imply that $[P, H] = 0$ and hence the momentum is conserved. In the classical argument one would normally say that the force $F(x) = -dU/dx = 0$ and Newton's second law then tells us that $dp/dt = F = 0$, leading to the same conclusion.

We encountered this situation for example in the section

about the 'free particle on a circle' of Chapter II.5 where we found that states were labeled by the quantized momentum $p = \hbar k$ (k-integer), being a conserved quantum number. So we chose a framework consisting of the energy and the momentum operator, with as *sampling space* just the momentum eigenvalues $-\infty \leq p \leq +\infty$. Here we see that if an underlying space-time symmetry, like translation invariance, is also present in the Hamiltonian, then indeed, the spectrum reflects that. But there is always a dual aspect. On the one hand the momentum P which is the conserved quantity, but on the other that very same P is the generator of the symmetry transformations being the translations. We have illustrated this general relationship in Figure II.6.1.

Rotations and angular momentum conservation. Let us now consider a more complicated example where symmetry tells us a lot about the spectrum, the case of the Hydrogen atom. The spectrum exhibited a large degeneracy which explained and depicted already in Chapter I.4 in Figure I.4.9. The states are labeled by three integer-valued quantum numbers: the energy related quantum number $n = 1, 2, \ldots$, the angular momentum quantum number $l = 0, 1, \ldots, n-1$ and the magnetic quantum number $-l \leq m \leq l$. In this problem we have a spherically symmetric electric force field centered at the nucleus in the origin. The energy consists of two parts, a kinetic part $\mathbf{p}^2/2m$ and a potential part $-k/|\mathbf{x}|$ and each part depends only on the length of the vectors and therefore is invariant under rotations. So we expect that the generators of rotations commute with the Hamiltonian and that they are therefore conserved, and somehow their sample spaces should be reflected in the labeling of the degenerate states with equal energy. Indeed, the generators of those rotations around the x, y, and z axes are the corresponding angular momentum observables/operators defined as a vector $\mathbf{L}$:

$$\mathbf{L} = \mathbf{X} \times \mathbf{P}\,.$$

Furthermore, the three components are conserved, as one can indeed show:

$$[H, L_i] = 0 \quad i = 1, 2, 3\,.$$

But now a further complication pops up: the conserved components of $\mathbf{L}$ do not commute among each other. We have:

$$[L_1, L_2] = i\hbar L_3\,, \quad \text{and cyclic permutations.} \qquad \text{(II.6.2)}$$

This algebra of real three-dimensional rotations, denoted as $so(3)$ happens to be identical to the by now familiar $su(2)$ Lie algebra. To describe the system we need to choose a framework $\mathcal{F}$, which means that we have to choose a subset of mutually commuting operators. Conventionally one chooses the following set: H, $\mathbf{L}^2 = L_1^2 + L_2^2 + L_3^2$ and L_3 with the eigenvalues:

$$\begin{aligned} H\,|\psi_{nlm}\rangle &= \frac{E_0}{n^2}\,|\psi_{nlm}\rangle\,; \\ \mathbf{L}^2\,|\psi_{nlm}\rangle &= \hbar^2 l(l+1)\,|\psi_{nlm}\rangle\,; \\ L_3\,|\psi_{nlm}\rangle &= \hbar m\,|\psi_{nlm}\rangle\,. \end{aligned} \qquad \text{(II.6.3)}$$

And as we mentioned before, for a fixed value of the principal quantum number n, there are in fact $2n^2$ degenerate states as a consequence of the symmetries that are present in the problem. The set of those states form a basis for all allowed states with an energy corresponding to that value of n. If we take $n = 3$, we should have $l = 0$, $l = 1$ and $l = 2$, but the symmetry algebra $so(3)$ given in (II.6.2) does not change the value of l, only the values of m from $-l$ to $+l$, which means that the rotational symmetry only accounts for the $(2l+1)$-fold degeneracy for each value of l. The conclusion therefore is that for $n = 3$, the spectrum consists of the three distinct *irreducible representations* of the rotation group (labeled by $l = 0, 1, 2$), see also Figure II.6.2. That suggests that there is may be more symmetry present in this problem, a topic we will return to shortly.

Let us make another observation here. In the choice of the framework we at once introduced the operator $\mathbf{L}^2$, which is

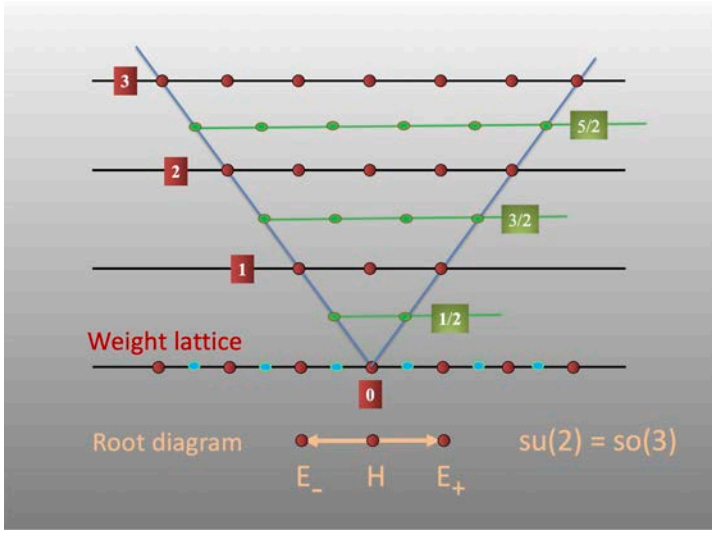


Figure II.6.2: *The representations of* $\mathfrak{su}(2) \simeq \mathfrak{so}(3)$. The group SU(2) has three generators that form the algebra $\mathfrak{su}(2)$. The root diagram has the diagonal L_3 which forms the Cartan subalgebra $\mathcal{H}$, while the arrows represent the raising and lowering operators $E_\pm \simeq L_\pm$. The weights of all (unitary) representations are on the weight lattice. We furthermore depicted the weight diagrams of various irreducible representations labeled by successively $l = 0, 1/2, 1, 3/2, \ldots$.

strictly speaking not part of the Lie algebra. It is a quadratic combination of generators that has the nice property that it commutes with all of the $\mathfrak{su}(2) \simeq \mathfrak{so}(3)$ generators: $[\mathbf{L}^2, \mathcal{A}] = 0$. Such invariant polynomials (also called Casimir operators or Racah invariants) play an important role in Lie algebra theory because you can use them to label or identify the inequivalent representations. And indeed, the eigenvalue $l(l+1)$ (or for that matter l) labels and distinguishes the infinitely many different (irreducible) representations of the algebra by $(2l+1) \times (2l+1)$ matrices.

Vectors and spinors. Let us return to the abstract algebra (II.6.2) of $\mathfrak{so}(3)$. We have mentioned that this algebra is identical to the algebra $\mathfrak{su}(2)$ generated by (a half times) the Pauli matrices X, Y, and Z. And this implies that the algebra not only has integer l representations, but also half-integral, so-called spinor, representations. And as you see these do not show up in the orbital angular momentum part, but in the part associated with the spin of a particle, which is a degree of freedom that is not present at the classical level. Actually saying that there is no classical equivalent is of course not correct. We have shown that the classical system underlying the spin-half, quantum degree of freedom, is just the classical two-state system of a bit or Ising spin. Not much 'rotational' about it and that is what is implied by saying that it has no classical analogue. But if you 'believe' the mathematics, the half-integral representations had to be there somewhere, and yes they showed up in the anomalous Zeeman-effect that brought Uhlenbeck and Goudsmit in 1925 to their bold conjecture of the 'intrinsic spin' of the electron, and 5 years later became a compulsory ingredient of any particle obeying the Dirac equation. This we discussed already in Chapter II.1.

So what we learned from these examples is that the Lie algebra $\mathfrak{so}(3)$ which happens to be the same as $\mathfrak{su}(2)$ has an infinity of inequivalent (unitary) representations labeled by an integer or half-integer quantum number $j = 0, \frac{1}{2}, 1, \ldots$ and that that representation can be realized by $(2j+1) \times (2j+1)$ hermitian matrices. There is a basic distinction between the integer and half-integer eigenvalue representations: physicists refer to the integer ones as *vector representations* and to the half-integer ones as *spinor representations*. In the hydrogen atom we saw all the representations showing up, in the discussions we had on the qubit we start off with a single spin one-half (doublet) representation, but as we mentioned before in the n-qubit space we have a much bigger symmetry group acting corresponding to $SU(2^n)$, which contains the product group of n individual $SU(2)$ as a subgroup.

An additional dynamical symmetry. Let us return to the spectrum of hydrogen and note that there is still something we haven't explained. The degeneracy observed at energy level n equals $2n^2$. It involves a degeneracy of *different* l representations, which cannot be accounted for by the

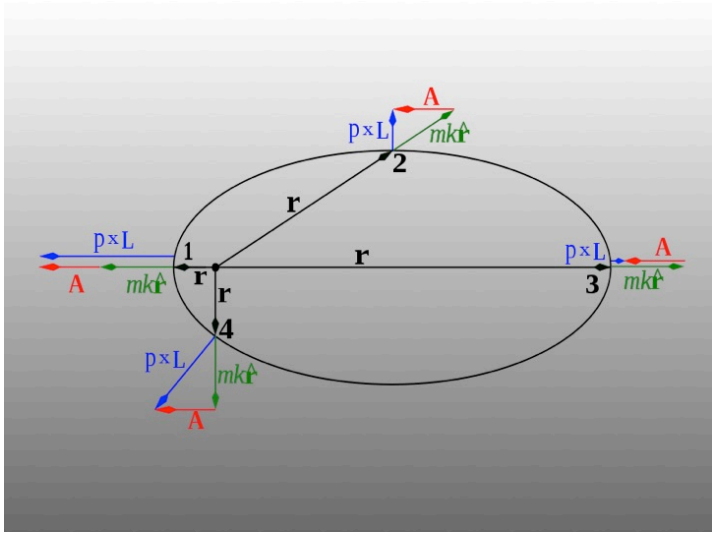


Figure II.6.3: *The Runge–Lenz vector,* $\mathbf{A} = \mathbf{P} \times \mathbf{L} - \mathrm{mk}\hat{\mathbf{r}}$ is an additional conserved quantity in the problem with a central -k/r potential, where the origin of the coordinate $\mathbf{r}$ is in one of the focal points. $\mathbf{A}$ points always parallel to the long axis of the ellipse in the direction of the 'perihelion.'

rotational symmetry. It could be 'accidental,' but that would be hard to believe if you have stayed with me so far. You would probably bet that it *must be* the consequence of yet another symmetry that we still have to disclose and that would make the whole picture even more striking.

Indeed, that symmetry is there, as there are in fact three more (independent) observables that commute with the Hamiltonian if it has a central 1/r potential. The presence of this symmetry is directly linked to the particular form of the interaction potential and is therefore called a *dynamical symmetry*. The generators form a vector just like the angular momentum and that vector is called the Runge–Lenz vector after its (re)discoverers.[3] This vector usually denoted by $\mathbf{A}$ is defined as:

$$\mathbf{A} = \mathbf{P} \times \mathbf{L} - \mathrm{mk}\hat{\mathbf{x}}\,. \tag{II.6.4}$$

[3]It has an interesting history with many rediscoveries going back to the early 18th century. Pauli was the first to use it to solve the hydrogen atom in an article from 1926.

We have constructed $\mathbf{A}$ at various points of a classical Newtonian elliptic orbit in Figure II.6.3, and we see that it is indeed a constant of the motion. Note that it takes some use of the 'like-rule' to get the orientation right and then you see that the vector is parallel to the long axis of the ellipse and points in the direction of the 'perihelion.' It is surprising that such a conserved vector-like quantity exists, but you expect on the quantum level to be responsible for the extra degeneracy with respect to the quantum number $\mathrm{l} = 0, 1, \ldots, \mathrm{n} - 1$.

That explains by the way that in the Newtonian theory the elliptic orbit is completely fixed in space, and moreover it also explains that this feature disappears if we add a correction term coming from Einstein's general theory of relativity. That term concerns a small $1/\mathrm{r}^3$ contribution, that breaks the symmetry and therefore the ellipse is no longer fixed in space and starts rotating in the plane of the orbit. This is the well-known 'perihelion precession' that was observed for the planet nearest to the sun *Mercury* already in the nineteenth century, and could indeed be accounted for by Einstein's theory. It illustrates the notion of an approximate symmetry it is not an exact symmetry but nevertheless teaches us about essential features of the system.

The full symmetry of the hydrogen atom

After all this struggling with vector products you may like to know what the total symmetry algebra of the hydrogen atom really is. This algebra is six-dimensional, and is indeed generated by the three $\mathbf{L}$ and the three $\mathbf{A}$ components. They form a closed algebra and it is in fact the algebra $so(4)$ of the rotations in four dimensions. So here we are, we set up a problem in three dimensions and now we get a spectrum exhibiting a manifest $so(4)$ symmetry. It underscores that the algebra has many representations and these may show up in all kinds of contexts

which have nothing whatsoever to do with a physical four-dimensional space. Here it surfaced because besides the rather evident spatial rotational symmetry of the problem, there turned out to be the additional, somewhat hidden *dynamical* symmetry (dynamical because it depends on the particular $1/r$ behavior of the potential and not on the underlying space). Including that symmetry allowed us to fully resolve the degeneracies in the hydrogen spectrum.

Raising and lowering operators. We see that we have chosen a consistent framework $\mathcal{F} = \{H, L^2, L_z\}$ to label the states. They are mutually commuting, but now you may ask what happened to the other symmetry operators – L_x and L_y for example – that commute with the Hamiltonian but *not* with L_z. We basically know what their meaning is as we showed before that they can be regrouped into raising and lowering operators that step up and down the different m values (within a single l representation). And similarly the components of the Runge–Lenz vector can be used to step up or down the value l of the total orbital angular momentum. So in this case these are operators that make steps not in energy but rather in other quantum numbers that label the degenerate states.

So if we go to the table on page 447, we see that a framework $\mathcal{F}$ typically involves a set of $\mathrm{rank}\ \mathcal{A}$ operators forming a so-called Cartan subalgebra $\mathcal{H}$ of $\mathcal{A}$. A Cartan subalgebra consists by definition of a maximal set of mutually commuting generators of $\mathbf{A}$. And indeed, the other generators in $\mathcal{A} - \mathcal{H}$ can be regrouped in a complete set of raising and lowering operators.

A full set of step and symmetry operators satisfying equation (II.5.21) is called the *spectrum generating algebra* for the obvious reason that they allow you to walk through the sample space, in principle finding all the energy eigenstates and their quantum numbers referring to a framework compatible with the energy operator.

Generating the spectrum (sample space). Let us assume that by some means we succeeded in constructing a complete set of step operators which bring you from one energy level to another, one could in principle imagine looking for the ground state(s) (the state(s) that are 'annihilated' by all the lowering operators) and then, using the spectrum generating algebra of all step and symmetry operators, to generate the whole spectrum of eigenstates of the Hamiltonian.

We have seen that symmetries, and in particular the maximal set of mutually commuting symmetry operators, yield the set of quantum numbers that allows us to label and distinguish a relevant basis for all states. And as the labels of such base states corresponds to eigenvalues of symmetry operators they are conserved in time. Therefore, in a general sense, such a maximal set allows us to 'name' the properties of the system, since 'names' are useful precisely because they do not change all the time. On the other hand if the system undergoes interactions, the properties may change and also then it is important to have a proper identification of property names or quantum numbers. For example, the interaction may excite the system and therefore basically act like a raising operator.

Symmetry algebra and symmetry group

So far we have talked about the observables Q_i that commute with the Hamiltonian. They are conserved and we have seen that they generate a symmetry. That means that acting with them gives an infinitesimal displacement corresponding to a tiny symmetry transformation. This applies of course only to the case of continuous symmetries. You might wonder what a *finite transformation* then would look like and how they are described. It is here that we have to move from the mathematical concept of a (Lie)-algebra to that of a Lie group. This question is briefly addressed in the *Math Excursion* on Vectors and Matrices on page

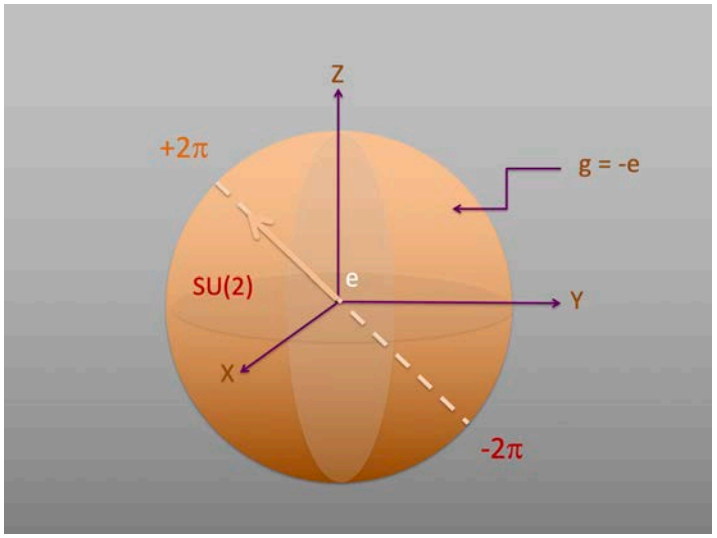


Figure II.6.4: *The group manifold of* $SU(2)$. $SU(2)$ can be represented as a solid three-dimensional ball with radius 2π. A point in that space corresponds with a rotation around the vector by an amount that corresponds to its length. All points on the surface are identified and represent minus the identity element: a rotation of about any axis by 2π yields an overall phase minus one.

635, and we have used in Chapter II.3 in the section on the Berry phase on page 347.

Exponentiation of the algebra. Let us return to the question of frame rotations for a qubit corresponding to a two-dimensional (complex) vector .

We considered the Z-frame and the X-frame and these frames are clearly related to each other by a finite rotation over an angle of 45^o around the y-axis (perpendicular to the $z-$ and $x-$axes). Let us make an angle rotation over an angle θ around the Y axis[4] and use the matrix version of the Euler identity:

$$e^{i\theta Y/2} = 1\cos\theta/2 + iY\sin\theta/2\,;$$
$$\Leftrightarrow R_y(\theta) = \begin{pmatrix} \cos\theta/2 & \sin\theta/2 \\ -\sin\theta/2 & \cos\theta/2 \end{pmatrix}. \qquad \text{(II.6.5)}$$

Let us apply this to see what it does with the basis vectors:

$$\begin{pmatrix} \cos\theta/2 & \sin\theta/2 \\ -\sin\theta/2 & \cos\theta/2 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos\theta/2 \\ -\sin\theta/2 \end{pmatrix}.$$

If we put $\theta = 90^o$, we get exactly the finite rotation of state $|+1\rangle$ to $|-\rangle$ as indicated in Figure II.2.1 where the frame choices are discussed and how these choices are related to the unitary group transformations we denoted as U in our discussion in Chapter II.1. We also know how to apply this transformation to the operators, We have to act from both sides for example:

$$Z \to R_y(\theta)\; Z\; R_y(-\theta) = -X\,,$$

where we have used the fact that $R_y(\theta/2)^\dagger = R_y(-\theta/2)$. This explicitly resolves a puzzle that you may have felt uneasy about. The algebra is three-dimensional with $X/2, Y/2$ and $Z/2$ as basis vectors, and indeed by rotating Z around the Y axis with $\theta = 90^o$ yields $-X$, exactly as you would expect, but applying the same transformation to the qubit rotates the two-dimensional 'vector' only over 45 degrees. How is that possible? Well to be precise the qubit is not a vector in the usual sense it is therefore that we introduced the term *spinor* exactly to make this distinction.

From the above considerations one may show that any finite $SU(2)$ group transformation can be parametrized as

$$g(\{\gamma^a\}) = e^{i\sum_a \gamma^a T_a} \quad \text{with } \{T_a\} = \{X/2, Y/2, Z/2\}.$$

Finite translations. For the translations one can do a similar exponentiation,

$$T(a) = e^{iaP}. \qquad \text{(II.6.6)}$$

[4] Here the factor a half comes back and becomes relevant. The parameter is θ, but the generator satisfying the su(2) commutation relations is Y/2, and therefore it looks like a rotation by $\theta/2$, but it is not.

which gives that on an operator which depends on X, and P we obtain after a finite translation by any amount a:

$$\begin{aligned} f(X,P) &\to T(a)f(X,P)T(-a) \\ &= e^{iaP}f(X,P)e^{-iaP} = f(X+a,P)\,. \end{aligned} \qquad \text{(II.6.7)}$$

In particular one has the property that $T(a)XT(-a) = X + a$., showing that the X operator has been shifted by a.

In the same vein you can show that if $[H,P] = 0$ and P is conserved. That also means that

$$T(a)HT(-a) = H\,,$$

which literally says that it leaves the Hamiltonian invariant, i.e. the translations are a symmetry of the Hamiltonian.

What I am trying to make plausible is that by 'exponentiating the algebra' we do get the corresponding group. Whereas the algebra describes infinitesimal transformations you need the group to do finite transformations. And whereas the algebra is a linear vector space, the group is some smooth curved manifold.

The group space or manifold of $SU(2)$. You can think of a group as a smooth manifold or space. For example, the group $U(1)$ is just a circle as we mentioned before. For the real space translations it is $\mathbb{R}^3$ because a finite translation in space is fixed by the three components of the displacement vector.

The group SU(2) is isomorphic to the three-sphere S^3 as we discussed in Chapter II.1 on page 254. So exponentiating the $su(2)$ algebra (note the use of lowercase) we get the $SU(2)$ group (in capitals). The $su(2)$ algebra has generators X, Y,and Z, and is therefore three-dimensional. The dimensionality of the algebra is the same as that of the group (manifold). The group $SU(2)$ has therefore three independent parameters, or coordinates. You can think

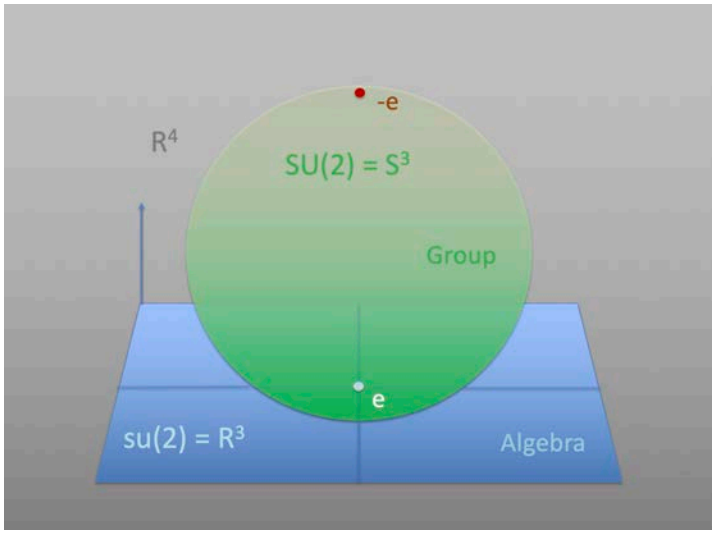


Figure II.6.5: *The group* $SU(2)$ *and its algebra* $su(2)$. $SU(2)$ can also be represented as a unit three-sphere S^3 embedded in R^4. The $su(2)$ algebra can then be thought of as the R^3 tangent space to the group manifold in the origin (the point corresponding to the trivial or unit element e).

of the algebra as the tangent (hyper) plane to the group manifold in the unit element e (corresponding to the trivial transformation). That plane has of course the same dimension but is a linear, flat space like $\mathbb{R}^n$. In Figure II.6.5 we give illustrated this relation between the $SU(2)$ group and the $su(2)$ algebra. If you stay near the unit element, a change in the tangent plane is almost as good as moving on the group manifold. It's like assuming that the Earth is flat, which is not such a bad approximation if you look on the scale of kilometers, but causes serious trouble if you start thinking in terms of thousands of kilometers! Thinking locally amounts to making a linear approximation, as for small $a \simeq \varepsilon$ we may write

$$T(\varepsilon) \simeq 1 + i\varepsilon P\,.$$

This terminology is that the algebra *generates* infinitesimal transformations. In short: thinking local acting global is bad, while thinking global and acting local is fine.

Gauge symmetries

We have argued that the equations that form the starting point for quantum fields are basically the same equations that one can write down for classical fields. Those classical fields change from being just functions on the configuration space to operator valued fields. And these then have to be quantized typically using canonical methods where the fields become like 'field coordinates' and their derivatives like 'field momenta'.

Electrodynamics revisited. Let us go back to the Schrödinger or better the Dirac equation in three plus one dimensions and ask how we could implement the interactions with the electromagnetic field. Somewhere in the equations there ought to appear terms that describe this interaction. Now we go through a beautiful argument where you will see how a number of rather peripheral remarks we have been making before all fall into place and yield a profound insight. That insight amounts to the fact that nature has a hidden symmetry and that imposing that symmetry completely fixes the precise form of the interactions (fundamental forces) between the elementary constituent particles.

I give the argument in relativistic notation, because that keeps things simple and elegant. The argument also holds true in non-relativistic situations. We want to use space-time vectors that have four components: for example instead of using the usual momentum vector $\mathbf{p}$ we switch to the four-momentum written p_μ where $\mu = 0, \ldots, 3$ and the time component of the four momentum is defined as $p_0 \equiv E/c$. Now if you look at the equations describing the interaction of charged particles with the electromagnetic field, then it turns out that you can get those interaction exactly right if you use a simple trick that goes by the name of 'minimal substitution'. It is a recipe that says: for a particle with a charge e replace everywhere the momentum p_μ by $p_\mu + eA_\mu$. The four vector $A_\mu = (V, \mathbf{A})$ are the electromagnetic potentials where V is the electrostatic or scalar potential and $\mathbf{A}$ as the vector potential.

These were introduced in the section on electrodynamics in Chapter I.1, together with the electromagnetic *field strength* $F_{\mu\nu}$:

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \,. \qquad \text{(II.6.8)}$$

The three spatial components F_{ij} correspond with the components of $\mathbf{B}$, and the space-time components F_{0i} correspond with the components of $\mathbf{E}$.

Gauge invariance. In Chapter I.1 we argued that there is some redundancy in keeping all the six components of the fields $\mathbf{E}$ and $\mathbf{B}$ and one could do with only the four components of the gauge potential A_μ. That is indeed the case but as a matter of fact even that doesn't eliminate all redundancy. In the formulation with the gauge potentials there is still some redundancy left, because we can make a transformation on the gauge potentials that leave the field strength F and thus the physical $\mathbf{E}$ and $\mathbf{B}$ fields invariant. This transformation is called a gauge transformation and involves a space-time dependent function $\Lambda(x,t)$:

$$A_\mu(x,t) \to A'_\mu(x,t) \equiv A_\mu + \partial_\mu \Lambda(x,t) \,. \qquad \text{(II.6.9)}$$

If you substitute the transformed field into (II.6.8), you immediately see that the extra terms cancel each other out, and that proves the invariance (and the efficiency of the relativistic notation).

This invariance is of another type than we have been discussing before, because the transformation depends on space-time. It is called a *local* transformation because by choosing the transformation you fix the amount by which you transform in every point independently, as long as it changes smoothly from one space-time point to the next. This means that we are effectively dealing with only three components for the gauge potential, because one may choose the gauge function in such a way as to 'gauge away' one of the components of the gauge potential. So,

why don't we get rid of it you may say, and strip the description of the electromagnetic field to the bare minimum. This is not so easy and you could say that keeping the redundancy is the price we pay for the transparency and compactness of the theory, and most importantly its linearity. This theory is beautiful like a peacock, with the exceptional property that it can fly as well! We are a bit like dealers in options when we talk about the field strengths which correspond to the invariant physical degrees of freedom, but these are in fact *derivatives* of the underlying potentials to which the particles couple.

Covariant derivative. The minimal substitution means that for charged particles we change the momentum operator to

$$P_\mu = -i\hbar\partial_\mu \to -i\hbar D_\mu \equiv -i\hbar(\partial_\mu + i\frac{e}{\hbar}A_\mu)\,; \qquad \text{(II.6.10)}$$

where e is of course the charge of the particle. In other words, the recipe is to replace the ordinary derivative ∂_μ by the covariant derivative D_μ.

We also remarked before that the Schrödinger or Dirac field is complex and therefore has a real and an imaginary part. And we furthermore made the point that there is always one overall phase that is unobservable and has no physical meaning therefore. Transforming that phase into another phase would not matter; it reshuffles the real and the imaginary parts of the wave function but the combination of the two has exactly the same content. Nevertheless, there is a phase symmetry because there is a phase transformation that leaves the physics invariant

$$\psi(x_\nu) \to \psi'(x_\nu) = e^{i\alpha}\psi(x_\nu)\,. \qquad \text{(II.6.11)}$$

Furthermore, the equations with the interaction term also are invariant under this phase transformation. This transformation is often called a *global*, meaning space-time independent gauge transformation.

Now we pose the interesting question whether these equations are also invariant under *local*, which means space-time dependent phase transformations:

$$\psi(x_\nu) \to \psi'(x_\nu) = e^{i\alpha(x_\nu)}\psi(x_\nu)\,.$$

On first inspection the answer is no, because the equations have derivatives that 'see' that space-time dependent phase factor and are going to make trouble about it because:

$$\partial_\mu\psi \to \partial_\mu\psi' = e^{i\alpha(x_\nu)}(\partial_\mu + i\partial_\mu\alpha(x_\nu))\psi\,;$$

and the transformed equation would be different because of this extra term involving the derivative of the space-time dependent phase. But wait a minute, what if we include the gauge potentials as we are supposed to do if we adopt the minimal substitution doctrine. Then we get:

$$D_\mu\psi \to (D_\mu\psi)' = e^{i\alpha(x_\nu)}(\partial_\mu + i\partial_\mu\alpha(x_\nu) - i\frac{e}{\hbar}A'_\mu)\psi\,.$$

Now please observe a tiny miracle, if we just substitute the expression (II.6.9) for gauge transformed A'_μ and make the judicious choice $\Lambda = (\hbar/e)\alpha$ then net the effect of the two transformations is zero and we get that the gauge covariant derivative transforms exactly as we want,

$$D_\mu\psi \to (D_\mu\psi)' = e^{i\alpha(x_\nu)}D_\mu\psi\,.$$

It transforms 'covariantly' just like the field ψ itself and therefore the complete theory involving also matter fields becomes gauge invariant. This result implies that the equations transform now simply by an overall local phase, which we can divide out and we have not changed anything.

We conclude that the complete system of Maxwell equations coupled to the Schrödinger or Dirac equations exhibits this local gauge invariance.

Gauge connection and parallel transport. The gauge invariant part of the electromagnetic field are the $\mathbf{E}$ and $\mathbf{B}$ fields, or the components of $F_{\mu\nu}$. But as we have been discussing already in the previous section on particle statistics and anyons there is a more subtle *non-local* quantity

that is gauge invariant, namely the Aharonov–Bohm phase factor or Wilson loop defined in equation (II.3.3).

If there is curvature (field strength) then the transport between point x_0 and x_1 becomes path dependent. The linear covariant equation:

$$D_\mu \psi(x) = 0\,;$$

has a general path dependent solution:

$$\psi(x_1) = e^{-i\frac{e}{\hbar}\int_{x_0}^{x_1} A_\mu dx^\mu} \psi(x_0)\,.$$

It looks quite daunting but think of it as just a phase factor, where the phase equals this integral of A_μ along the path, which is after all just a real number. This expression tells you precisely what *parallel transport* means: it tells you how the electromagnetic phase changes if you move in position space. And the covariant derivative in (II.6.10), is the infinitesimal version of that. The first term with the derivative generates a translation, while the second generates the phase transformation. This also connects with the entries in the table on page 447, the exponent is a phase factor corresponding to a group element of the group $U(1)$ which is just a circle. And A_μ is the *connection one-form* which takes a value in the Lie algebra which is just the phase itself. $U(1)$ is one-dimensional group, and it is generated by a 'one by one hermitian matrix': in other words a real number.

The other point is that this ties in perfectly with our earlier observations in the previous section concerning the Aharonov–Bohm phase factor, as a means of measuring the magnetic flux up to multiples of the basic flux quantum $2\pi\hbar/q$. The remarkable aspect is that the path may entirely lie in a region where the electric and magnetic fields themselves are zero, yet the closed loop measures a nontrivial and gauge invariant quantity. It measures a topological aspect of the theory.

We finally recall the other application of the parallel transport notion as a way to measure some Hamiltonian landscape by means of the so-called Berry phase, as we discussed in Chapter II.3. There, the notion of parallel transport was used to detect 'curvature' or 'field strength' differences between a flat and curved surface.

Charge conservation. We have emphasized over and again that one of the reasons why symmetry is important is that it corresponds to conservation laws. In fact there is a basic theorem by the German 19th century mathematician Emmy Noether that to any one parameter continuous symmetry there is an associated conserved 'charge.' Local symmetries include the corresponding global symmetry and one therefore expects that the gauge symmetries will also correspond to conserved quantities. For the electromagnetic gauge symmetry that is – not surprisingly – the local conservation of electric charge.

A rather direct proof of this was already presented in the subsection on gauge invariance on page 33 of Chapter I.1. Recall that the interaction of the field with an external current gives a contribution to the Lagrangian density of $A^\mu j_\mu$. So if we make the gauge transformation we get only one extra term which equals $+ie(\partial^\mu \Lambda/\hbar) j_\mu$ in the Lagrangian density, because the current itself is assumed to be gauge invariant. Invariance of the theory requires this extra term after integration over space-time to vanish. This in turn requires that the current has to satisfy $\partial^\mu j_\mu = 0$ which amounts to the local conservation of charge. This equation tells you that the change of the charge in that volume exactly equals the current going through the surface bounding that volume. This is the relativistic form of what we in general call a *continuity equation* which is a local conservation law indeed.

Turning arguments around. A question that you might have raised is whether we could have turned the arguments around and have said: let us *impose* this invariance under local transformations on the Dirac or Schrödinger

equations, what do we have to do? The answer would have been: you have to introduce a gauge potential A_μ that transforms in such a way that it absorbs the troublesome extra term coming from the derivative. So introducing gauge fields is a necessary consequence of imposing local gauge invariance.

It was through arguments along these lines that in 1954 the physicists Chen Ning Yang and Robert L. Mills discovered the structure of *non-abelian gauge theories* that form the backbone of the acclaimed Standard Model.

Non-abelian gauge theories

In this section we go through the steps that brought Yang and Mills to what must have been an incredible *eureka* moment: the discovery of non-abelian gauge theories.

Think of our familiar qubit as a column vector with two complex entries, but now we make it into a complex two-component spinor or doublet field, which we denote it by $\psi(x_\nu)$ and we have the derivative ∂_μ which can act on it. Next we want to make a field theory for ψ that is locally gauge invariant. The first thing is to ask what invariance there is under constant or global transformations. Well, it is not just a single phase but it can be any unitary frame rotation U as we discussed for example in the *Math Excursion* on page 635 of Part III. Such rotations correspond to elements of the group $SU(2)$, and we learned that any element of the can be written as the exponent of an element of the $su(2)$ algebra which is a linear combination of the Pauli matrices:

$$U(\gamma) = e^{iC} \quad \text{with} \quad C = \gamma_1 X + \gamma_2 Y + \gamma_3 Z \equiv \boldsymbol{\gamma}\cdot\mathbf{T}\,.$$

By construction C is hermitian $(C^\dagger = C)$ and U therefore unitary $(U^\dagger = U^{-1})$. Now we want to repeat the exercise we did for the phase factor with this matrix valued 'phase'.

Gauge covariant derivative. First we observe that the derivative has still no problem with the constant complex rotation by which we mean that the three components of $\boldsymbol{\gamma}$ are constant. But what if the parameters become space-time dependent, if we write $\boldsymbol{\gamma} = \boldsymbol{\gamma}(x_\nu)$, and look what happens with at the two-component derivative if we transform $\psi(x_\nu) \to U(x_\nu)\psi(x_\nu)$

$$\begin{aligned}(D_\mu\psi) &= (1\partial_\mu + iqA_\mu)\psi \to \\ (D_\mu\psi)' &= (1\partial_\mu + iqA'_\mu)U\psi \\ &= U(1\partial_\mu + U^{-1}\partial_\mu U + iqU^{-1}A'_\mu U)\psi \\ &= U\,(D_\mu\psi)\,. \end{aligned} \tag{II.6.12}$$

In the first line we should now think of the covariant derivative as a matrix where the derivative is multiplied with the unit matrix and A is some matrix with a structure we are about to determine. The strength of the coupling between the A and ψ fields is given by the charge q. In the intermediate line we have inserted the trivial factor $U\,U^{-1} = 1$ in front, in order to obtain the expression in the desired form, which appears in the bottom line. But that expression only holds if the gauge field A has the interesting structure which is more or less dictated by the derivative term:

$$U^{-1}\partial_\mu\, U = U^{-1}(\partial_\mu\boldsymbol{\gamma})\cdot\mathbf{T}\,U\,.$$

Because the factors U, U^{-1} and $\mathbf{T}$ are matrices they do not commute and one cannot just change the order in which they appear in an expression.

Lie algebra valued gauge fields. Apparently this derivative brings down the Lie algebra element and takes the derivative of that, and the result of this gets rotated by the U factors around it. The upshot is that this non-abelian gauge field has to be an element of that same Lie algebra so:

$$A_\mu = \mathbf{A}_\mu\cdot\mathbf{T}$$

and it has to transforms like:

$$A_\mu \to A'_\mu = UA_\mu U^{-1} + \frac{i}{q}(\partial_\mu\, U)\,U^{-1}\,.$$

For the case at hand the conclusion is now clear, the gauge field itself has to be an element of the Lie algebra in this case $\mathfrak{su}(2)$, and has to transform like a connection. The is algebra is three-dimensional as it has three independent generators, and consequently there are three independent gauge fields needed, which represent three different gauge particles.

Principle fiber bundles. The appropriate mathematical setting of gauge theories is that of *fiber bundles*, as we discussed already in the section on the 'Physics of geometry' on page 78 of Chapter I.2. These bundles are defined as a triple $\{E, M, \pi\}$ corresponding to a *bundle space* E, a *base manifold* M (which would be our space-time manifold) and a *gauge (or structure) group* G. The dimension of E equals the sum of the dimensions of M and G. And the space E looks locally like a tensor product $M \otimes G$, but can be different globally, in which case we speak of a nontrivial bundle. Given is a *projection* π from E onto M, and the inverse of that projection at a point $x_\mu \in M$ gives you the *fiber* above that point which is a copy of (isomorphic to) G. Choosing a smooth *section*, meaning that you choose a particular group element out of each fiber, produces an explicit form of a gauge covariant derivative on M. Gauge transformations are related to the changing of sections of the bundle.

This setting allows you to naturally define topologically nontrivial gauge field configurations that can be characterized by topological invariants like the *Chern classes*. Deep results relevant for physics were obtained. For example, a variety of the so-called index theorems, like the *Atiyah–Singer index theorem*, that links the topological invariant of the gauge field configuration to the net number of left- versus right-handed solutions of the zero-mass Dirac equation coupled to that (background) field. Interestingly the Yang–Mills equations were not considered before they appeared in the physics literature, and only afterwards became a major mathematical topic in the 1970s.

Once more the Standard Model. We mentioned that the number of gauge particles is equal to the dimension of the Lie algebra, which is just the number of independent parameters or generators. But the argument does not depend on the particulars and basically holds for any gauge group, including the groups $U(1)$, $SU(2)$, and $SU(3)$ that appear in the Standard Model. The weak and electromagnetic interactions have the gauge group $SU(2) \times U(1)$, where the charged $W^\pm$ bosons correspond to the raising and lowering operators $T_\pm$, while the photon and the neutral Z boson are linear combinations of the neutral W^0 boson and the Y boson associated with the $U(1)$ factor of the gauge group. The three W bosons correspond thus with the three-dimensional (iso) spin 1 representation in Figure II.6.2, while the fermionic quarks and lepton fields form doublets corresponding to the (iso) spin-$1/2$ representation.

Colors and Flavors. Quantum Chromodynamics (QCD), the theory for the strong interactions, has gauge group $SU(3)$, which has dimension eight. The eight gluons correspond with the weights of the root diagram (including two zero weights in the center) as shown in Figure II.6.6. In this figure we have also marked the color (anti-)triplet representations corresponding to the weights of the (anti-) quark fields.[5]

At this point you may experience a *deja vu* moment, because Figure I.4.33 in Chapter I.4 flashed back in your mind which indeed looks very similar to Figure II.6.6. Yes, true, but it actually refers to a very different context. There we were talking about the *flavor symmetry*, the classification scheme discovered by Gell-man and Zweig. It is indeed also an $SU(3)$ symmetry, and it also applies to the quarks but on the other hand it is a very different type

[5] The gluon circles carry a quark and anti-quark color, and we have given the anti-quarks the anti- or better complementary color in the figure. In Figure I.4.36 the gluons are also bicolored but there both the quark and antiquark have the same color but have arrows in the opposite direction.

of $\mathrm{SU}(3)$ symmetry. Firstly, it is not a gauged symmetry, but instead an approximate global or rigid symmetry, so there are no gauge particles associated with it. And as the quarks of different flavors have different masses it is indeed only an approximate symmetry, because the particle states are not really degenerate. Our knowledge at this point suggests that this symmetry is accidental, and once you accept that it is only approximate you may as well declare that there is a $\mathrm{SU}(4)$ or even $\mathrm{SU}(6)$ flavor symmetry. This would be the case if you in addition take the charm, top and bottom quark flavors along. Anyway, the physics related to these two $\mathrm{SU}(3)$ groups is entirely different: the flavor symmetry is manifest in the spectrum of observed particles, as the figure in Chapter 4 shows. The mesons for example belong to an octet and these are free particles. The color property of particles is hidden because of the confinement phenomenon which only allows color neutral or singlet states to be free particles. This made it so hard to uncover the color symmetry in the first place.

Color singlets. The singlet property has to do with constructing colorless combinations of quarks (and gluons). This requires that we look in the possible multi-quark spectrum for those combinations which have that property. Here I recall the fact that multi-particle states are described by so-called tensor products of single particle Hilbert spaces. The single (anti-)quark color states form a color (anti-)triplet representation denoted as 3 and $\bar{3}$ respectively. The tensor products can be split up again in irreducible components or representations. Like for example:

$$\begin{aligned} 3 \times \bar{3} &= 1 + 8 \\ 3 \times 3 &= \bar{3} + 6 \\ 3 \times 3 \times 3 &= 1 + 8 + 8 + 10\,. \end{aligned} \qquad \text{(II.6.13)}$$

The dimension of the tensor product space is the product of the dimensions of the two factors. The weights of the tensor product states are obtained by adding the weights of the individual representations. This you may verify in the SU(3) weight space of Figure II.6.6. What is clear from equation (II.6.13) is that the simplest ways to make a color singlet '1' representation is by combining a quark and an anti-quark, making a *meson*, or making a particular combination of three quarks making a *baryon*.

Is Einstein gravity a gauge theory? So we have found that the gauge symmetry principle underlies the particular way the force carrying particles appear in nature. Does this trick then also work for the gravitational force you may wonder. Yes indeed, it does! One interesting way to interpret the Einstein theory is actually to look at it as a gauged version of the combined *local* Lorentz and translation groups, usually referred to as the *Poincaré group*. So in this perspective the Einstein equations are an expression of a local Poincaré symmetry.

Kaluza–Klein theory. You could also argue the opposite way and say that the E and B fields, the field strengths of electromagnetism, correspond to electromagnetic 'curvatures' of some internal space that is defined in every point in space-time. Yet another way to understand it is to say that space-time has in fact extra spatial dimensions, which have particular geometries corresponding to circles, spheres or group manifolds for that matter. These compact extra spaces are then squeezed to zero size, by a procedure called 'dimensional reduction' or 'dimensional compactification.' This remarkable idea in fact goes back to the early days of general relativity where Theodor Kaluza and Oskar Klein proposed to unify electromagnetism and gravity in a five-dimensional theory using this symmetry principle.

The proper mathematical setting for the classical versions of gauge theories is that of *fiber bundles* with some Lie group G or representation thereof as fibers, as we introduced them in the section on 'The physics of geometry' on page 78 of Chapter I.2. These geometric structures attracted the attention of the physicists only long after the not so geometric Maxwell, Einstein and Yang–Mills equations

were written down. In fact the formalisms were developed to a large extent independently in physics and mathematics.

Non-abelian field strengths. You might complain that I am choking my highly esteemed readers with math, but to my defence I would argue that we have exposed some of the core ideas of modern physics, in only a few pages , and even without too much cheating! In fact the 1954 paper of Yang and Mills is just a short article that appeared in the *Physical Review Letters* (PRL) journal, and its influence is inversely proportional to the length of the paper. There is an ironic aspect to that paper, since the authors in fact proposed that this non-abelian gauge theory should describe the 'pions' as these particles were at that time believed to mediate the strong nuclear interactions. This idea didn't work out at all, and so these beautiful equations went into the 'fridge,' and it took about 15 years before they were taken out again and found their true vocation in the Standard Model as we have described it.[6] It is one of those rare occasions where the elegance and beauty of an idea make it irresistible and fortunately also inescapable, so one just had to wait for it to find its proper place.

You might object by noting that the Kaluza–Klein idea of dimensional compactification apparently has *not* properly landed, in spite of being attractive and elegant as it 'produces' gauge fields with the correct interactions. The K–K approach returned as a necessary ingredient of string theory, but nevertheless has not yet found its true vocation, and I am afraid it has to spend some more time in the 'fridge.' Science is patient and even if an idea clearly 'does not work,' it is extremely hard to put stickers stating 'Consume before date indicated on the bottom' on ideas.

[6] A hallmark of great institutions is not only that they attract extremely gifted people, but also that they are the guardians of research fields, keeping alive a collective memory of failed attempts and almost forgotten, unsolved problems; of all that ended up in the 'fridge of ideas' so to speak.

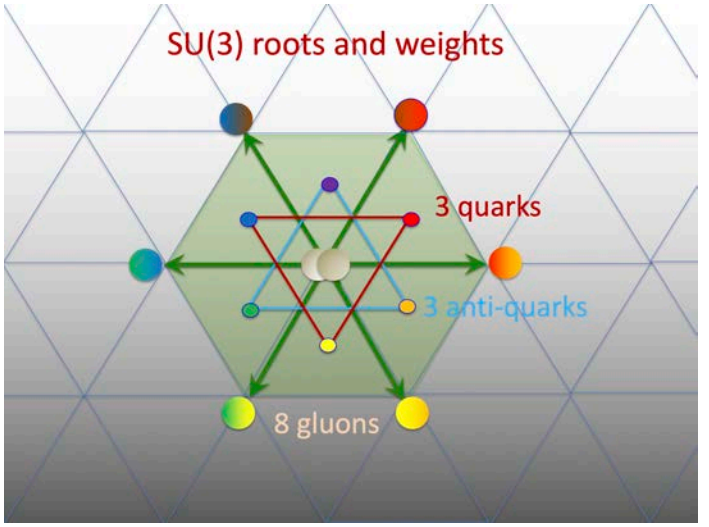


Figure II.6.6: $\mathrm{SU}(3)$ *roots and weights.* In this figure we represented the root diagram of $\mathrm{SU}(3)$. with the 6 non-zero roots given by the green arrows. The gluons form the 8 representation corresponding to the six non-zero roots and the two in the center, marked by the bi-colored circles. Then there are the triplet (3) and the anti-triplet $(\bar{3})$ representations corresponding to the three colored (anti-)quarks.

.

The Yang-Mills equations

So are we done? No, not quite, we have to check one other thing: what will happen to the analogue of the Maxwell equations for the gauge fields? And what happens to the electric and magnetic fields, so nicely encoded in the field strength $F_{\mu\nu}$, if we go non-abelian? Two remarks are to be made, (i) as F is linear in the gauge field it also will live in the Lie algebra and should therefore simply transform as $F \to F' = UFU^{-1}$ and (ii) this is only achieved if the definition of F for the non-abelian theories is generalized in a logical and elegant way to:

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + iq[A_\mu, A_\nu]\,; \qquad \text{(II.6.14)}$$

logical and elegant because the commutator is antisymmetric in the indices and also keeps you in the Lie algebra. An equivalent, more covariantly looking definition is to

say that $F_{\mu\nu} = -\frac{i}{q}[D_\mu, D_\nu]$. The extra commutator term in the field strength has huge physical consequences it turns out.

Clearly the definition of the non-abelian electric and magnetic fields are nonlinear in the potentials, and this means that the Yang–Mills equations, which are the generalizations of the Maxwell equations to the non-abelian case, are nonlinear as well. The Yang–Mills equations really are the dynamical expression of non-abelian gauge symmetry. These equations take the following form:

$$D^\mu F_{\mu\nu} = \partial^\mu F_{\mu\nu} + ig[A^\mu, F_{\mu\nu}] = 0\,.$$

They are strongly nonlinear indeed, in the first place because the definition of the field strength is non-linear in A, and secondly because of the presence of the commutator term of A with F in the equation itself.

Symmetry dictates the structure of interactions. The non-linearities mean that the theory is self-interacting right from the start. Whereas photons don't see each other, gluons do, as we already showed in Figures I.4.36 and I.4.37. We have reproduced the latter here to take a closer look at how it connects to the more detailed description of non-abelian gauge theories we have given.

The local Lagrangian density is a Lorentz invariant expression for non-abelian gauge fields (gluons) coupled to Dirac fermions (quarks) and looks deceivingly simple:

$$\mathcal{L}(x_\mu) = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\bar{\psi}\not{D}\psi\,, \qquad \text{(II.6.15)}$$

with $\not{D} = \gamma^\mu D_\mu$ the Lorentz invariant Dirac operator as it works on a four-component Dirac field $\psi(x_\mu)$. In Figure II.6.7 we see two interaction vertices: on the left we see a self-interaction of the gauge field corresponding to the third order term in A from the F^2 term in the Lagrangian and on the right we see the gauge field interact with the Dirac field corresponding to the cubic interaction term from the covariant Dirac operator in the Lagrangian. There is a lot

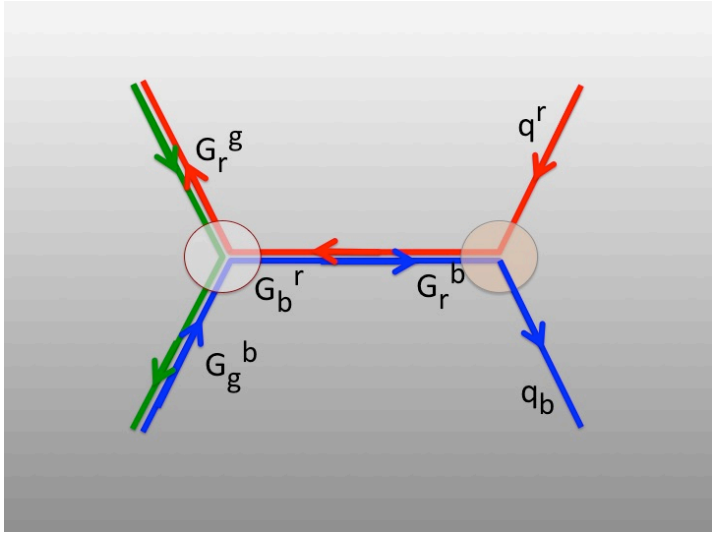


Figure II.6.7: *Color-flow diagram in QCD.* A nice way to visualize the interactions in QCD. Quarks carry a single color line, while gluons carry two (different) lines. In the vertices the color charge is conserved, so, the colors and arrows have to match. The upper index goes into the vertex, the lower index goes out.

of index gymnastics hidden in the notation however. This becomes evident if we for example write out the latter term in glorious detail. It looks quite horrendous:

$$i\frac{q}{\hbar c}\bar{\psi}(x_\nu)^i_a\gamma^\mu_{ij}A(x_\nu)^\gamma_\mu T_\gamma{}^{a\,b}\psi(x_\nu)^j_b\,. \qquad \text{(II.6.16)}$$

There are a few remarks to make with respect to this intricate expression:

(i) the interaction is local as all fields depend on the same space-time point x_ν;

(ii) all fields carry a space-time index that tells you how they transform under Lorentz transformations, and a gauge index that tells you how it transforms under gauge transformations;

(iii) the Dirac fields carry two indices, a space-time spinor index i with $i,j = 1,\ldots,4$, and a 'color' index a with $a,b = 1,\ldots,n$ with n the dimension of the color representation (n=3 for QCD);

(iv) the four gamma matrices carry a space-time (vector) index μ and each of them is a matrix in spinor space and has therefore two spinor indices i, j;

(v) the gauge field has a space-time index μ and a gauge group index γ with $\gamma = 1, \ldots, \dim \mathcal{A}$ ($\dim \mathcal{A} = 8$ for QCD);

(vi) the representation matrices or generators T carry a gauge group label γ and each of them is a matrix in the representation space, thus with two indices a, b;

(vii) all indices are pairwise contracted, and thus have to be summed over. This amounts to making invariant inner products in the spaces the indices refer to. In a sense the expression is therefore extremely simple because once you know what the symbols stand for, there is a strict logic which tells you where to put the various indices. It is dictated by the requirement of invariance of the interaction under independent changes of basis in either space-time, or spinor space, or in the Lie algebra or group representation spaces;

(viii) It is these delicate balancing act of indices that is for example reflected in the way the 'color 'lines in the diagram of Figure II.6.7 are strictly continuous through the vertices.

Some people might say that it is ugly to exhibit all these indices, while others say that that is exactly what makes the very beauty of the construction manifest. The ultra compact notation of equation (II.6.15) demonstrates how effective the symbolic notation is that the physicists have developed over the years. The expression (II.6.16) in contrast shows very explicitly how a particle in fact lives in many spaces simultaneously, all with their own indices and metrics. All of us agree that to do real calculations you have to go all the way down into this index jungle, it is a must, a *conditio sine qua non!* And once you realize in addition that this is only the lowest order interaction diagram you can imagine that it takes a fully dedicated PhD researcher to complete a single higher order calculation of some physically relevant process that is measured in an accelerator. Such calculations involve hundreds or even thousands of diagrams to be added to get the full probability amplitude for the process. The actual execution of such calculations involves nowadays high-level AI in large scale computing efforts and it is thanks to the rigorous underlying symmetry structures that these calculations can be automated to such a large extent.

Self-interactions and the confinement problem. Free fields are sometimes not as free as one would think. And this in turn makes perturbative approximations dangerous, which basically means that you start with setting the coupling strength q to zero, and then take only low orders of q into account. The problem is that if a field is self-interacting the theory becomes nonlinear and may end up in a phase which is entirely different from what you naively would expect. The relevant or observable degrees of freedom can be very different from the degrees of freedom you started out with. For example the enigmatic problem of *quark confinement* can be traced back to the self-interacting nature of the gluons. Free quarks have never been observed, because they are doomed. They have to spend their whole life as a pair, or a *ménage a trois* but always confined within a hadron.

Understanding and proving these quantum confinement properties of Yang–Mills theories from first principles is still an open question and is one of the Millennium problems in mathematics. It is a problem that attracts the minds of brilliant mathematicians and theorists because it is a very well-defined problem. The starting point is a familiar object called a non-abelian gauge theory, or a principal fiber bundle with a compact structure group. The quantum problem to be solved is: prove the conjectured confinement property of the 'color-electric' fields. That this property holds has been demonstrated by numerous computer simulations of the theory, where the theory is formulated on a discrete space-time lattice but that amounts basically to a study of the strong coupling (large q) limit of the theory. This is basically a perturbative approach in $1/q$. And in

that limit the theory does confine, but to settle the question one has to prove that there is not a phase transition between the strong and weak coupling regime.

From experience – think for example of Fermat's last conjecture – we know that such conjectures can linger around for centuries before they finally get turned into a theorem. A humble observation is that making the conjecture already can make you famous. So the least we can say is that we are exploring deep waters!

The conclusion is that the principle of local gauge invariance provided a valuable clue to the construction and understanding of the fundamental equations underlying the Standard Model. ■ ■

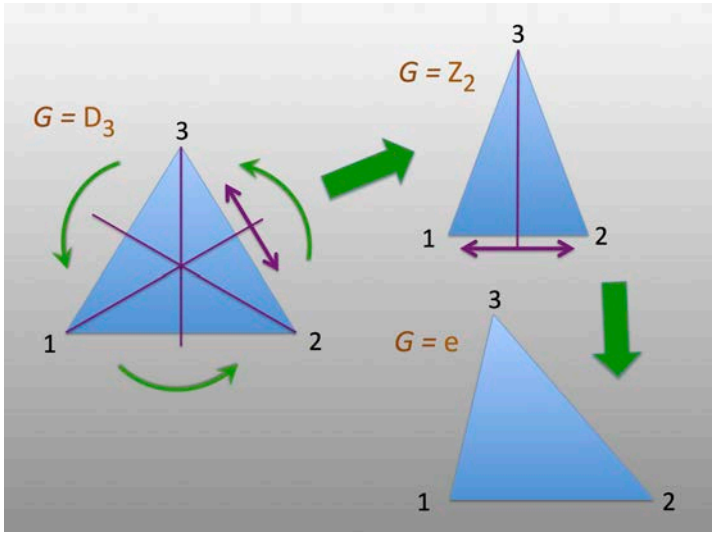


Figure II.6.8: *The breaking of symmetries.* Breaking symmetry by deformation of an object. The D_3 symmetry of the equilateral triangle (with six elements) gets broken to a Z_2 of isosceles triangle (with two elements), which subsequently gets broken to the trivial group for an arbitrary triangle.

The symmetry breaking paradigm

Having argued that symmetry principles play an important role in modern-day physics, the same can be said about the concept of symmetry breaking which has found many beautiful and surprising applications in basic high-energy physics as well as in many branches of condensed matter and molecular physics. Where symmetry unifies states and makes them degenerate, it is the breaking of symmetries which creates non-uniformity and diversity. We are going to explore some typical cases which illustrate the power of this quantessential idea.

Symmetry breaking in objects. It is paradoxical that I first let you suffer by talking so extensively about how beautiful symmetries are, and then immediately after confront you with how to break them. It is like a small child building a beautiful tower from woodblocks and then destroying it while screaming and dancing around it. Apparently there is some thrill in the act of destruction! Let us look for similar thrills, and first go back to the 'symmetries of objects', like an equilateral triangle, a circle or a sphere, and then it is not hard to imagine how to break the symmetry.

For example you could squeeze the object one way or another as to reduce its symmetry. You could do it step wise like in Figure II.6.8, where you first go from an equilateral to an isosceles triangle, and then to a generic one. In that case you first pass from the discrete group D_3 with six elements (3 rotations and 3 reflections) to the group Z_2 of two elements (the identity element and a reflection), and in the second step you end up with no symmetry at all: you are left with only the identity element. Breaking has the property that the residual symmetry group after breaking is just a subgroup of the original symmetry group.

If you squeeze a ball top down, you typically get an ellipsoid, where the symmetry is reduced to rotations around the vertical axis only, and a reflection symmetry through the horizontal plane and vertical planes through the center.

If you then make it into a standing egg shape, you lose the reflection property in the horizontal plane but still keep the vertical symmetry axis, and so on. By the way this makes you wonder why eggs have the shape they have. Why not celebrate the perfection of life in perfect spheres? One reason that has been given is that egg-shaped objects do not roll away, if you put them on the table and push them away they tend to 'boomerang' in a little circle. 'They like to stay near their starting point!' I hear my mother say. And maybe the biology of how to lay an egg – to push it out by contraction – plays a role as well in the optimal egg design. What came first, the egg or the design? This is not even a 'chicken or egg' question, instead, this is an 'egg or egg' question. Anyway, more a topic in evolutionary biology than in quantum physics I fear, so it is better to leave it to the *cloaca* experts. The shapes created by symmetry breaking are more and more diverse and need more and more parameters to specify. In that sense their information content and therefore entropy increases. And many will say that with that their beauty increases as well.

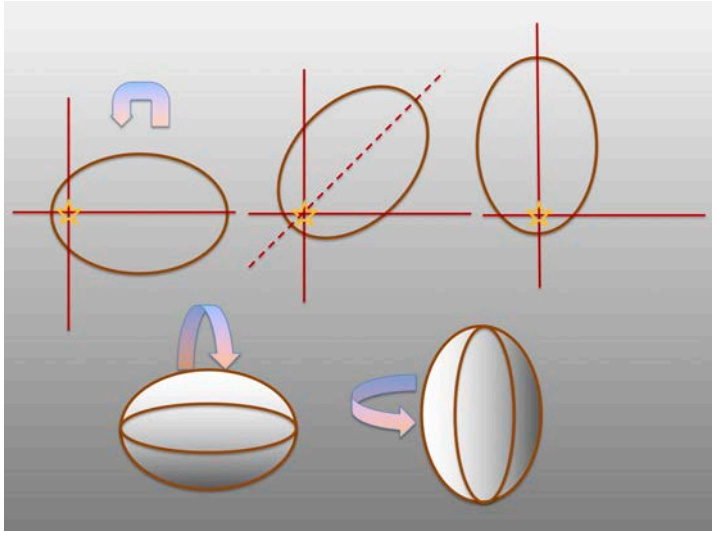


Figure II.6.9: *Action of rotational symmetry on an elliptic orbit solution.* The Newtonian Earth–Sun system has spherical symmetry but that symmetry is not manifest in a particular solution, like for example an elliptic orbit. The symmetry transforms different equal energy solutions into each other.

Symmetry breaking by solutions of equations. The next step up is to talk about the symmetry of equations, and the first question that comes to mind is what do the solutions of equations with symmetries look like? Do they indeed manifestly exhibit the symmetries of the equations? The answer is clearly: No! Think of our nice Newtonian example again. The great step forward was exactly to discover and understand that the planetary orbits are *not* circles or even epicycles, but conic sections, ellipses, parabolas and hyperbolas. So, where did the spherical symmetry of the gravitational field around the sun go, which is so clearly present in the equations? Why and where does the immaculate perfection of the heavenly spheres get lost?

A little thinking yields the answer: the symmetry is still there. But the symmetry transformations act on the space of solutions. What they do is that given a particular solution, and acting with a symmetry operator on it, it will in general generate a different solution. The symmetries map solutions onto each other, and as they keep the equations fixed, they transform solutions with equal energy into each other. With the rotations that is quite obvious, as we have illustrated with the elliptic orbits of the spherically symmetric Newtonian sun-earth system in Figure II.6.9. It turns out that the Runge–Lenz symmetry changes the eccentricity of the elliptic orbit and that is not so obvious. It is in this sense that you may say that most particular solutions break the symmetry of the equation, and the symmetry acts in the space of solutions. It creates a subspace of degenerate solutions in the space of all solutions. That space gets 'stratified' according to its energy values and solution shapes.

This brings us in fact close to the observations we have made with respect to the role symmetries play in quantum theory, labeling the degenerate states but also moving (stepping) between them. They walk you through the

degenerate subspace of the total sample space of your favorite framework.

Symmetry breaking in the atom. Symmetry breaking is an important concept. What does symmetry breaking look like in a quantum setting? Imagine that we have a symmetry, then we could make that symmetry visible by 'breaking it'. In other words by adding a term to the Hamiltonian that explicitly breaks the symmetry. For example we put an atom in a magnetic field say along the z-direction, then there will be an extra term in the Hamiltonian proportional to $\mathrm{L_z}$ and the magnitude B of the magnetic field. Now the three-dimensional rotational symmetry is broken to rotations around the z-axis only. The consequence is that the energy levels which were at first degenerate and therefore hard to distinguish will now split up proportional to the value of their magnetic quantum number m. This is the famous splitting first observed by Pieter Zeeman we discussed in Chapter I.4. This is an example of *explicit symmetry breaking* where we change the Hamiltonian. But also in quantum theory we can have the phenomenon of *spontaneous symmetry breaking* which refers to a situation where we change external parameters of the system – say the temperature or a coupling – such that the Hamiltonian itself does not change and still has all the symmetries, but it is the ground state that changes to one in which the symmetry is broken.

Low energy modes. This brings us to a follow-up question: what happens if the ground state is not invariant and does not respect all the symmetries? In other words, what if the ground state breaks the symmetry? Well, by what we argued above, it will then necessarily be the case that that ground state is not unique and itself degenerate. If that ground state breaks a *continuous* symmetry, we will have a *continuous* set of equivalent ground states. And what that means is intuitively quite clear: the system can easily move from one ground state to one nearby and it would cost basically no energy.

Figure II.6.10: *Long-range orientational order.* The collective of wheat plants is in a state that exhibits a long-range order. By growing out of spherical seeds, the original rotational symmetry is broken.

Saying it yet differently, the generators of the symmetries that are broken create 'zero (energy) modes' of the system. This is an important physical signature of broken symmetry: the appearance of low energy modes in the system that are easy to excite. And if we talk about (relativistic) field theory where the energy includes also the mass, our observation asserts that there will be massless particles around. Such particles are called Goldstone particles or modes, after the MIT physicist Jeffrey Goldstone who discovered the mechanism. Ideally these modes are exactly massless, but there can be additional effects that give those particles a mass. However, that mass should be small compared with the scale of the interaction energy that caused the breaking.

Think of wheat seeds, if we assume them to be spherical, spreading them on a field gives a 'ground' or better 'down to earth' state that is rotationally invariant, which means to say that we can rotate each of the seeds by the same amount and nothing will change. Now we wait

Figure II.6.11: *Wheat waves as low energy excitations.* The wheat field at rest shows that long-range order typical for a broken symmetry. It has low energy excitations. These are the 'wheat waves' that propagate easily and can already be excited by a gentle breeze.

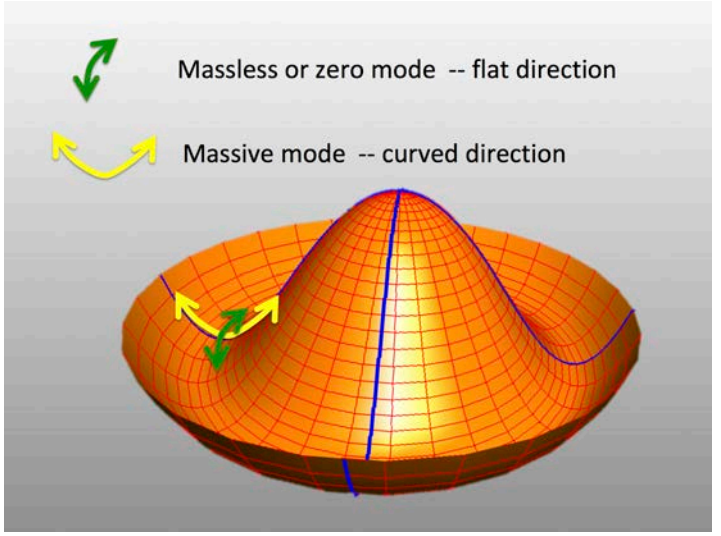


Figure II.6.12: *Breaking of global symmetry.* The breaking of a $\mathrm{U}(1)$ global symmetry leading to a 'Mexican hat' potential. The minimum is not unique but there is a continuum of ground states forming a circle. The breaking leads to one massless and one massive mode as indicated in the figure.

a month or more, and the seeds turn into plants nicely growing up, all beautifully lined up vertically, so the ground state has changed to a 'field' in a completely ordered state that certainly is a state of broken symmetry. There is a spontaneous, average length which is non-zero, and furthermore a long-range vertical orientational order in the system which breaks the original spherical symmetry (see Figure II.6.10).

Now where is the zero mode? Those modes correspond to what you get if a light breeze goes over the field and you see gentle plane waves traverse the wheat plants (see Figure II.6.11). It is a low energy collective mode that originates in the broken symmetry of the ground state. Amusing and playful for sure, but we better take it seriously because there are many examples of this so-called Goldstone mechanism, from spin waves or *magnons* in magnets, to the appearance of three nuclear particles known as *pions*, $\pi^{\pm}$ and π_0 we have mentioned in Chapter I.4.

Chiral symmetry breaking. A famous application of the symmetry breaking concept is provided by the three pion particles ($\pi_{\pm}\mathrm{and}\pi_0$). The interpretation is that they are the Goldstone particles associated with what is called *chiral symmetry breaking*. It refers to an ingenious scenario proposed by Japanese/American physicist Yoichiro Nambu, who indeed received the Physics Nobel prize in 2008, for – I quote – 'the discovery of the mechanism of spontaneous broken symmetry in subatomic physics.' The scenario starts with massless *up* (u) and *down* (d) quarks. These are described by massless Dirac equations, but the massless Dirac equation can be split into two non-interacting pieces, the right (R) and left (L) polarized components. Said differently, it is precisely the mass term in the equations that couples the left to the right polarized components. If you look at the tables of the standard model in Figure I.4.35, you see that there is the horizontal so-called *isospin* symmetry between u and d quarks. This means that the massless equations have an $\mathrm{SU}(2)_{\mathrm{L}}$ symmetry

transforming u_L and d_L into each other (so they form an isospin one-half representation), and an $SU(2)_R$ symmetry transforming the right-handed components u_R and d_R into each other. So at this stage the model has a six--dimensional $G = SU(2)_L \otimes SU(2)_R$ symmetry. This is called the *chiral symmetry group*, which is a global symmetry. Nambu suggested that a quark anti-quark condensate forms spontaneously, so that the particular diagonal combination of fields u and d becomes the order parameter and acquires a vacuum expectation value:

$$\langle \phi \rangle = \langle (\bar{u}_L u_R + \bar{d}_L d_R) \rangle = f_\pi \neq 0$$

Now this condensate breaks the symmetry G, but not completely. What is left you can see from the condensate, namely, if we *simultaneously* transform left *and* right then the condensate is invariant. This in turn tells us that from the six generators a particular 'vector like', 'left plus right' $SU(2)$ subgroup survives, while the rest, the three 'left minus right' generators, will be broken. These give rise to three Goldstone particles with exactly the quantum numbers that correspond to the three pion particles. The fact that these particles in the end do have a relatively small mass is accounted for by the fact that the masses of the quarks were not quite zero to start off with.

The breaking systematics. In the chapters on condensed matter physics we will return to this topic of symmetry breaking in the context of many body physics. The general picture boils down to a situation where the theory has a continuous symmetry group G of dimension $\dim G$, and some field gets a non-zero ground state expectation value. That particular vacuum state is only invariant under a residual symmetry group $K \subset G$ which is a subgroup of G. Then there will be $\dim G - \dim K$ broken symmetries and therefore the same number of Goldstone modes. The field that acquires the non-zero expectation value in the ground state and breaks the symmetry is called an *order parameter* field. The nomenclature is that the broken state is the state in which everything is neatly lined up some way and therefore exhibits 'order', where order is defined as the presence of long-range correlations in the medium.

Ferromagnetism. As an example, think of a metal where all the nuclear magnetic spins in the absence of an external field are pointing in random directions in the medium, and therefore there is no over-all magnetization, and no macroscopic direction of the magnetic field is discernible. If one then lowers the temperature below what is called the *Curie temperature*, the thermal energy gets so small that the weak interaction between the tiny magnets starts to become dominant and the spins minimize their energy by lining up and thereby 'spontaneously' make a magnet. So, by cooling down a metal spontaneous magnetization occurs and conversely, by heating up a magnet to high temperature it will lose its magnetization and the symmetry will be restored. Spontaneous magnetization serves as the prototype of spontaneous symmetry breaking in a many-body system. And indeed, the low energy modes are just the spin waves which are easy to excite in a magnetized medium.

Topological defects. In Volume III of the book we will address another crucial aspect of symmetry breaking, which is the appearance of what are called *topological defects*. Defects are collective excitations which are usually 'heavy' and not so simple to excite, but once they exist they are equally hard to get rid of. A dramatic instance you are all familiar with from watching the news is the phenomenon of tornadoes or vortices in liquids. In that case there is a ground state that is symmetric if there is no wind, but when a wind starts blowing there is at once everywhere at any given point in space, we have a local vector pointing in the direction of the wind. On the surface of the earth, we can think of the non-zero two-dimensional vector field representing the wind as an order parameter. As a consequence of some 'massive' obstacles it may happen that somewhere a pair of vortices with opposite vorticities is created, and once these get well separated, they are highly stable objects. As a matter of fact, you cannot destroy single vortices by locally disturbing them, you have to wait

till their energy gets dissipated, for example by causing a lot of damage. There are many examples of remarkably stable collective excitations in all kinds of fields of science and technology that can be thought of as topological defects that originate in a state of broken symmetry.

Hidden gauge symmetries: the Higgs particle. So far, we have only looked at rigid or global symmetries: we considered transformations that were the same at any point in space and we found the remarkable directly observable phenomena of low energy Goldstone modes and high-energy defects as hallmarks of their breaking. The next question that naturally arises in field theory is what happens if we somehow 'break' a local gauge symmetry? You may think of the $U(1)$ gauge symmetry of electrodynamics, or the $SU(2)$ gauge symmetry of the weak interactions. Again, this may happen *spontaneously,* meaning to say that the system of equations still has the full symmetry, but that the solution, in particular the ground state, does not. The first question to answer is whether this can be done at all. Is it possible to maintain the local gauge symmetry and yet have a ground state in which some field acquires a non-zero expectation value? The answer turns out to be 'approximately yes.' A first example was exhibited by Landau and Ginsberg in their effective description of superconductivity. Later it was understood and explained in full detail in the modern theories of Bardeen, Cooper and Schrieffer, and later Anderson, about which we have more to say in Chapter III.3.

The Brout–Englert–Higgs (BEH) mechanism

A beautiful example of the spontaneous breaking of a non-abelian gauge symmetry is the Brout–Englert–Higgs mechanism, accounting for the heavy mass of the weak force mediating $W^{\pm}$ and Z^0 particles in the weak and electromagnetic interactions, and more indirectly for the existence of the Higgs particle. Let me illustrate how that comes about in a simpler model due to Sheldon Glashow, without going into much detail.

Breaking in an $SU(2)$ model. Let us consider an $SU(2)$ (or $SO(3)$) gauge theory coupled to a 'matter' field that transforms like a triplet or iso-vector under the gauge group. This means that we should think of the gauge field as $A^a_\mu T_a$, where the T_a are now the three 3×3 matrices generating the $SO(3)$ symmetry. It has three gauge particles (like the W-bosons we discussed before) because the group is three-dimensional. The 'matter' field $\phi(x)$, is a triplet of space-time scalar fields, that transform like a 3-dimensional 'iso-vector' under the $SO(3)$ gauge group. In the quantum context the field $\phi(x)$ would therefore describe three types of scalar particles. Let us now assume that this field $\phi(x)$, or rather its square which is gauge invariant, develops a constant vacuum expectation value $\langle|\phi|^2\rangle \neq 0$. So a condensate forms. The situation is similar to the magnets we just discussed, but now we think of it happening in some internal space where the force field is active, and where the ϕ field describes an *iso-vector* degree of freedom at every point in space.

As long as the vacuum expectation value vanishes the symmetry is not broken, but if the iso-vector is non-zero, and chooses some fixed direction it is like a wheat field and the non-zero vector field is only invariant under rotations around the axis in the direction in which the nonzero iso-vector points, corresponding to an $SO(2)$ subgroup of dimension one. So, we expect there to be two massless Goldstone particles, like in the case we discussed before. But now in addition we have the gauge fields that are coupled to this iso-vector through a covariant derivative. The question is then what the effect of the vacuum expectation for the scalar field has on the gauge fields. The resulting mechanism is powerful and quite universal.

To see what happens we write for the iso-vector (and think of it as a three-component column-vector) in the 'broken'

phase:

$$\phi(\mathbf{x}) = \phi_0\,\hat{\mathbf{e}}_3 + \delta\phi(\mathbf{x})\,,$$

where ϕ_0 is the constant non-zero vacuum expectation value pointing in the third direction of iso-space and the delta describes the field fluctuations around that ground state value. Now the interactions are generated by the covariant derivative:

$$D_\mu\phi = (\mathbf{1}\partial_\mu + iqA^a_\mu\,T_a)\;\phi\,,$$

where the T_a are the three generators of rotations in iso-space. At this point the crucial observation is that there are two components of the gauge field that 'see' or sense the vacuum value, while the third component does not because it linked to the generator of the residual symmetry which leaves the condensate unchanged. The Lagrangian density $\mathcal{L}$ of this theory contains a term proportional to $(D_\mu\phi)^2$. Of interest here is only what the effect is of the constant ϕ_0 in the Hamiltonian. When you work out the interaction between the gauge field and the vacuum term you discover that it leads to a quadratic term proportional to $|\phi_0|^2$ of the form:

$$\Delta L = |\phi_0|^2((A^1)^2 + (A^2)^2)\,;$$

and this is exactly what a mass term for the two components of the gauge field would look like. Apparently we have generated a mass for two of the three force carrying particles, a mass proportional to the non-zero expectation value ϕ_0. So, we end up with one massless force component (A^3), which is long-range like the photon, and two massive force particles A^1 and A^2. The latter two can be recombined in the components $A^\pm$ which are charged with respect to the massless A^3 field. Because of their mass these fields mediate a short-range interaction described by a Yukawa potential as we explained in Chapter I.4. They would be the lookalikes of the $W^\pm$ particles. What we just described amounts to a simplified analogue of the Brout–Englert–Higgs mechanism in the Standard Model, which indeed explains the masses of the W and Z bosons mediating the weak interactions, and the photon remaining massless.

Searching for the Higgs. The remaining question is where does the celebrated Higgs particle reside in this scenario? I have not yet mentioned it. To understand its origin, we have to do some counting of the degrees of freedom of the particles before and after the condensate forms.

Let us start with a massless force mediator like the photon. In Chapter I.1 we showed that the photon field A_μ has two transversal polarization states orthogonal to its propagation direction. It is important to know that this transversality has everything to do with the fact that the photon is massless and, as we have argued before, it is the gauge invariance that effectively removes one degree of freedom from the three-component 'vector' potential. It is indeed the gauge invariance that – so to speak – protects the masslessness of a gauge particle like the photon. To get massive it would need the extra (longitudinal) component which is just not there, *basta*!

To continue our counting exercise, each component of the iso-vector field ϕ_i represents one field degree of freedom. independent of whether it is massive or massless. Suppose we take it to be a massive field, then after breaking, we create two massless Goldstone degrees of freedom while the third iso-component remains massive. Now comes the magic of the Higgs mechanism: the massless modes of the ϕ field get 'eaten' by the corresponding gauge particles, who become *stante-pede* massive after this exquisite meal. Because a massive vector field needs three polarization states, it has two transversal components like the photon, but also a longitudinal component, which the massless photon does not have. So, the upshot of the exercise is crystal clear: if we 'break' a gauge symmetry then the forces in the unbroken group stay unchanged but the force mediating particles that correspond to the broken generators, become massive and therefore short-range. And they become massive by absorbing the would-be Goldstone modes, which consequently disappear from the spectrum. There are no massless Goldstone particles but instead we have two massive vector bosons!

And now, to finally answer the question that got us into all this counting in the first place, where is that Higgs particle? The answer can only be that that particle corresponds exactly to the single leftover massive degree of freedom, the third component of that iso-vector Φ we started off with. So it is not the massless Goldstone degree of freedom that signals the breaking in this gauge symmetry setting, but the smoking gun is a neutral (it does not couple to surviving photon-like particle) massive scalar particle. What we learn is that the Higgs particle is not the condensate which gives the force carriers mass, but rather the quantized wave that rides on top of that condensate! It is a bit like having a transition from vapor to liquid water, which after the transition allows for waves propagate on the water surface. The degrees of freedom that acquire mass are the ones that have to wade through the water which makes them feel heavy indeed. The Higgs particle is the necessary a witness without alibi of this beautiful but intricate mechanism. The discovery of this unique feature that vindicates the BEH mechanism, a backbone of the Standard Model, by the ATLAS and CMS collaborations at CERN in 2012 was therefore a landmark discovery.

The mixing of weak and electromagnetic interactions. In the example above we have looked at the breaking of an $SO(3)$ symmetry by a non-zero vacuum expectation value of an iso-vector or triplet field ϕ , giving rise to masses for two of the three gauge fields. This is not quite the way the symmetry breaking works in the Standard Model. In the sector of the weak and electromagnetic interactions we have a gauge group $SU(2) \times U(1)$ involving the three gauge fields $W^{\pm}$ and W^0 for the $SU(2)$, and a gauge field Y for the $U(1)$ factor. This group is broken to a residual $U(1)_\gamma$, corresponding to the massless photon. This can be achieved by a non-vanishing expectation value for a scalar field that transforms like a doublet under the $SU(2)$ and is also charged with respect to the $U(1)_Y$ field. The net effect is that one is left with three massive gauge particles: the $W^{\pm}$ and the neutral Z_0 boson, which is a linear combination of the W_3 and Y fields. The other, orthogonal linear combination of those two neutral fields corresponds to the photon. This intricate mixing of symmetries shows reminds us of the fact that nature not always celebrates ultimate simplicity.

A symmetry not broken, but hidden. The above account of the BEH mechanism can be criticized on valid grounds. It may even be called misleading. I used this narrative for pedagogical reasons, because it borrows some of the vocabulary of the global symmetry breaking scenario. But a deeper fact is that the vacuum expectation value as I discussed it is gauge dependent. Because of the local gauge invariance, I can locally transform that vacuum vector in any direction I want, so the analogy with the phenomena of magnetization where that direction is directly observable and fixed is wrong. The good way to talk about the BEH mechanism is to say the invariant square of the covariant derivative acquires a vacuum expectation value, which directly translates into the mass terms for the vector particles. In other words there is a way of talking about this so-called breaking in a gauge invariant way. But then we have arrived at a *contradictio in terminis*, because if the mechanism can be cast in gauge invariant terms, then the gauge symmetry cannot be broken! Indeed! This is the reason that we rather speak of a *hidden symmetry*, the gauge invariance is still present, but is no longer manifest in the physics (the mass degeneracy), it is hidden. It is better to say that the gauge symmetry is not broken at all but realized in a different way in this physical model. This point of view is strongly supported by the technical fact that there is not necessarily a real phase transition between the hidden and manifest symmetric (confining) phase of the system.

Other forms of symmetry. We have in passing already referred to other symmetry types then the ones we have been considering here.

An important extension of space-time symmetries to so-called *supersymmetries* was a remarkable achievement.

The related super-algebras are not of the Lie algebra type, because they also involve fermionic generators that obey anti-commutators. If these extended symmetries are made local by gauging them, you need to introduce a spin-3/2 *gravitino* as the super partner of the graviton. As the names suggest these symmetries play a vital role in super string theory and super gravity theories and we commented on them at the end of Chapter I.4. The experimental program at the Large Hadron Collider at CERN has been searching for the lightest super particle that should exist in any supersymmetric theory with broken supersymmetry. And as we have not run into any superpartner of any particle in the Standard Model we have to assume that supersymmetry should be broken already at a high energy well above 1TeV.

Later, a remarkable class of algebra's were discovered, these are called infinite dimensional Lie algebras that are also known as *Kac–Moody algebras*. They have found interesting applications in two-dimensional physics both in string and condensed matter theory. It is a very high level of symmetry. After what we have said before one expects in this case there to be an infinite number of conservation laws, which almost tantamount to saying that models in which they feature, in spite of being very nonlinear are basically exactly solvable.

Finally, there is a class of symmetries related to what we called topological phases in matter, which are called *Hopf algebras* or *quantum groups*. The remarkable aspect of their application in two-dimensional physics is that their representations describe both the ordinary excitations, and the topological defects and their dyonic mixtures called anyons. These correspond to the exotic particles we briefly described towards the end of the previous chapter.

A detailed discussion of the symmetries we just mentioned is beyond the scope of this book, but we mention them to emphasize the richness of the symmetry concept in mathematical physics.

Symmetry concepts and terminology

We have explored many aspects of the notion of symmetry in this chapter. First we searched for the observables $\mathrm{Q_i}$ that commute with the Hamiltonian. These correspond with *conserved quantities* and form some Lie-algebra including the Hamiltonian H, which is then called the *symmetry algebra* $\mathcal{Q}$. The states of the system at some fixed value of the energy will form a degenerate set that corresponds to certain representations of the symmetry algebra. The degenerate states can be labeled by the eigenvalues of some mutually commuting subset of the symmetry generators, forming a so-called *Cartan subalgebra* $\mathcal{H}$ of the symmetry algebra. The choice of Cartan subalgebra corresponds to choosing a framework $\mathcal{F}$. The other symmetry operators that are not in the Cartan subalgebra can be combined into *raising and lowering operators* that walk you through the sample space of the chosen framework. In the following table we have summarized the correspondence between the physical and mathematical concepts underlying the notion of symmetry.

Math: *Group theory* $\vdots \quad \vdots$	$\supset$ **Continuous symmetries** $\subset$	**Physics:** *Quantum theory* Hilbert space of states Algebra of observables
Lie algebra $\mathcal{A}$ $\dim \mathcal{A} = d$	Observables Hermitian Commutator algebra Infinitesimal transformations Invariant polynomials (Casimirs)	$\{A_i\} = \{A, B, \ldots\}, \quad i = 1 \ldots, d$ $A^\dagger = A$ $[A, B] = iC$ $\Delta_A \lvert\psi\rangle = iA \lvert\psi\rangle$ $\{C_k\} \ \ (k = 1, \ldots, \mathrm{rank}\, \mathcal{A}) \ \ [C_k, \mathcal{A}] = 0$
Cartan subalgebra $\mathcal{H}$ $\dim \mathcal{H} = \mathrm{rank}\, \mathcal{A} = r$	$\mathcal{H} \subset \mathcal{A} \Leftrightarrow$ Framework $\mathcal{F}$ Mutually commuting (= Abelian) Labels basis states of representation N Weight vectors $\{\boldsymbol{\lambda}_m\}$	$\{H_i\} \quad i = 1, \ldots, r \leftrightarrow \mathbf{H}$ $[H_i, H_j] = 0$ $\lvert\psi\rangle_N = \sum_m c_m \lvert\{\boldsymbol{\lambda}_m\}\rangle_N$ $\mathbf{H} \lvert\{\boldsymbol{\lambda}_m\}\rangle_N = \boldsymbol{\lambda}_m \lvert\{\boldsymbol{\lambda}_m\}\rangle_N \ , \ m = 1, \ldots, N$
Cartan-Weil basis: $\mathcal{A} = \{H_i, E_{\pm\alpha_k}\}$ Root system $\{\pm\boldsymbol{\alpha}_k\}$ of $\mathcal{A}$ in $\mathbb{R}^d$	Raising and lowering operators	$E_{\pm\alpha_k} \quad k = 1, \ldots, (d-r)/2$ $[\mathbf{H}, E_{\pm\alpha_k}] = \pm\boldsymbol{\alpha}_k E_{\pm\alpha_k}$ $[E_{\pm\alpha_k}, E_{\pm\alpha_k}] = \pm\boldsymbol{\alpha}_k \cdot \mathbf{H}$
Symmetry algebra $\mathcal{Q}$ $\dim \mathcal{H} \leq \dim \mathcal{Q} \leq \dim \mathcal{A}$	Subalgebra $\mathcal{Q} \subset \mathcal{A}$ All Q_i commute with Hamiltonian H_0 $Q_i \sim$ conserved quantities $Q_i \sim$ generate symmetry transformations Time independent labeling of states	$\{Q_i\}$ $[Q_i, H_0] = 0$ $\dfrac{dQ_i}{dt} = 0$ if $H_0 \in \mathcal{H} \Rightarrow \mathcal{H} \subset \mathcal{Q} \Rightarrow \ \{\lambda_i\} \subset \{q_i\}$
Lie group $\mathcal{G}$ $\dim \mathcal{G} = \dim \mathcal{A}$	Unitary reps Transformation group on Hilbert space $\mathcal{H}_0$ Finite transformations: $\mathcal{G} \simeq e^{i\mathcal{A}}$ Group space coordinates	$U^\dagger = U^{-1}$ $\lvert\psi\rangle \to \lvert\psi\rangle' = U\lvert\psi\rangle$ $A \to A' = UAU^\dagger$ $g = e^{i\sum \gamma^i A_i}$ $\{\gamma_i\}$

On symmetries:

- *The Theory of Groups and Quantum Mechanics*
  Hermann Weyl
  (Reprint of 1931 Edition)
  Martino Fine Books (2014)
- *Symmetries in Fundamental Physics*
  Kurt Sundermeyer
  Springer (2013)
- *Symmetries and Conservation Laws in Particle Physics:*
  An Introduction to Group Theory for Particle Physicists
  Stephen Haywood
  Imperial College Press (2010)
- *Concepts of Elementary Particle Physics*
  Michael E. Peskin
  Oxford University Press (2019)
- *Aspects of Symmetry: Selected Erice Lectures*
  Sidney Coleman
  Cambridge University Press (1985)

On non-abelian gauge theories:

- *Quantum Field Theory and the Standard Model*
  Matthew D. Schwartz
  Cambridge University Press(2013)
- *Gauge Theories in Particle Physics*
  A Practical Introduction, Volume 2: Non-Abelian Gauge Theories: QCD and The Electroweak Theory
  Ian J.R. Aitchison (Author) and Anthony J.G. Hey
  CRC Press (2013)

Indices

Subject index Volume II

Name index Volume II

SANDER BAIS

Hierarchies: the emergence of diversity

The Quantessence of Reality

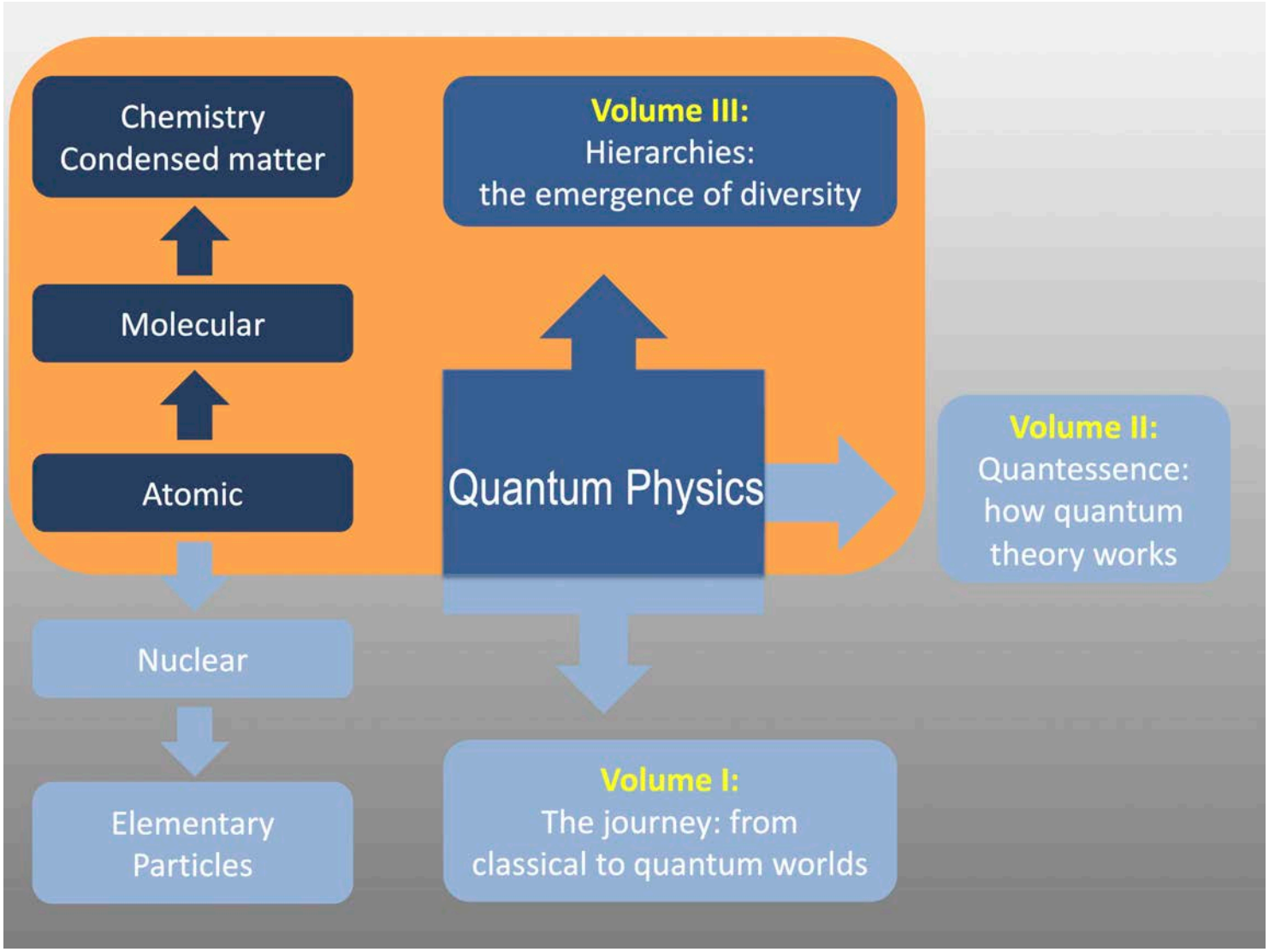
Amsterdam
University
Press

Hierarchies: The Emergence of Diversity

In this volume we describe the structural hierarchy underlying the visible world starting from atoms all the way up to the splendid diversity of condensend matter phases and basic chemistry. The properties of ordinary, liquid, and quasi crystals are explained as well as the complex behavior of the electron collective as we see it in different media varying from insulators to superconductors and from semiconductors to quantum Hall systems. To familiarize the reader with the necessary basic mathematics, a custom-made chapter Mathematical Excursions is included. This final volume closes with the chapter Chronologies, Ideas and People which provides tables of important quantum discoveries and the Nobel prizes awarded for them.

Volume III

Hierarchies:
the emergence of diversity


Chemistry
Condensed matter
Molecular
Atomic
Nuclear
Elementary
Particles
Volume III:
Hierarchies:
the emergence of diversity
Quantum Physics
Volume II:
Quantessence:
how quantum
theory works
Volume I:
The journey: from
classical to quantum worlds

Contents

II Quantessence: how quantum theory works

Contents **239**

III Hierarchies: the emergence of diversity

Contents **461**

Chapter III.1

The structural hierarchy of matter

Collective behavior and the emergence of complexity

> The behavior of large and complex aggregates of elementary particles, it turns out, is not to be understood in terms of simple extrapolation of the properties of a few particles. Instead, at each level of complexity entirely new properties appear and the understanding of the new behaviors requires research, I think, as fundamental in its nature as any other.
>
> P.W. Anderson in *More is different* (1972)

If we start from a large number of simple constituent particles which have simple interactions with each other, the collective of such particles may well exhibit a rich structural diversity and complexity. If we manage to identify the relevant collective degrees of freedom in the macroscopic system, then another simplicity may be regained, however. And relevance is what counts. This approach may reveal a hidden order and allow for an effective description of the apparent chaos and complexity in a limited number of variables.

Lost individuality. Let us start with a human analogy. Think of a couple, if they never talk to each other or seem to communicate, you'll treat them as separate individuals. You think of their 'relation' as a minor perturbation on their existence as individuals. However, if they are close and their relationship is a kind of symbiotic, you will treat the pair as a single entity: *they* are nice or crazy, or stupid. Their individuality is neither visible nor relevant it seems, what becomes relevant are the properties of the couple and these may be totally different from those of the individual.

Constituents and their interactions. The two cases represent two different regimes, which you might call *weakly* or *strongly* coupled. In the strongly coupled regime the next question is how the couples interact with each other, because that will have decisive implications for the collective behavior of a large crowd of people. To understand collective behavior one has to have some insight in the different aggregation levels below, in what the relevant agents at various levels are and how they interact. Are they individuals, couples, families or communities?

The differences in social organization between bees, ants, dolphins and humans can only be partially traced back to the difference in their specific species-linked features (for example the way their genetic information is passed over to the next generation) but to a large extent the social hierarchies they form depend on the nature of their interactions.

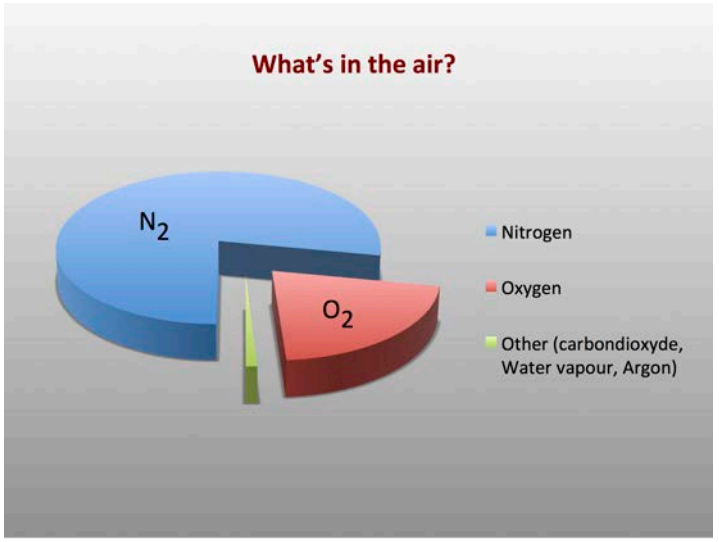


Figure III.1.2: *What's up in the air?* Air is a mixture of chemicals, and note that the nitrogen and oxygen components consist of the diatomic molecules N_2 and O_2. These atoms – like people – prefer to pair up somehow.

External parameters. Yet, there are still other important factors that play a role. Given the properties of the relevant constituents and their interactions, there may be different ways society becomes organized. In general it will also depend on external 'environmental' factors and dynamics. Revolutions may take place where a society reorganizes itself rather drastically. Depending on the external parameters it may go through 'tipping points.' A society may choose to adopt a new constitution, thereby redefining the basic set of behavioral rules. As external observer you usually don't directly observe the constitution, rather what happens as a consequence of it. What you may see is that the collective behavior changes drastically. And you may wonder whether they changed the constitution or whether the reason was a financial crisis for example.

What 's (up) in the air? Similar questions arise in physics if one wants to understand the binding of atoms into molecules or into macroscopic media like solids, liquids or gases. An everyday example is ordinary air: it is predominantly made up of the simple elements nitrogen and oxygen, and minor fractions of carbon, hydrogen and argon. But, in fact air is a mixture of chemical composites, since the nitrogen and oxygen have paired up (but for example not tripled up) while the others appear in composites like water vapor and carbon dioxide. Argon is the only element in the mixture perfectly happy on its own, an ideal *Einzelgänger* precisely because its electrons fill an entire shell of orbits, and this makes the atom inert, literally like a closed quantum shell.

From physics to chemistry to biology to... Here we enter the vast domain of chemistry, and condensed forms of matter in general, including the modern material sciences, biochemistry and molecular biology. These fields of science concern mesoscopic or macroscopic systems, which are characterized by a specific hierarchy of aggregation levels. The actual structural outcome may drastically change depending on external factors like density, temperature and pressure. The system may go through a so-called *phase transition*, where it reconstitutes itself in a tumultuous way before ending up in a new stable lowest energy ground state that may be drastically different from the state it started out from. We all know that water molecules can manifest themselves collectively in many radically different guises such as vapor, liquid and ice, but also in alternative structures like raindrops, hail and a huge morphological variety of snowflakes.

Emergent behavior. You can compare the ground state of a medium with what the constitution is for a human society. You do not observe it directly, only through the emergent behavior of the collective excitations it supports. The constitution is manifest in the way the society functions, or dysfunctions for that matter. It is the great variety in ways that matter has organized itself, which made it very hard to figure out what the constituents were in the first place. In this quest for ever more fundamental building blocks unrestrained reductionism reigned as we witnessed in Chapter I.4. To provide a broader context for the main subject of

this book we will in the remainder of this chapter highlight some representative examples of *structural hierarchies* of ever increasing complexity. And these emergent hierarchies are in some way or another the collective expression of the underlying quantum principles.

The ascent of matter

Cosmic evolution. The hierarchy of structures found in nature is quite universal. If we think bottom up, we start with the stable constituent particles of the Standard Model as depicted in Figure I.4.35, in particular the up and down quarks, and the electron. From a history of science perspective, working bottom up is anti-historical in the sense that the most basic constituents are the ones that have been discovered most recently, while many of the chemical compounds have been known for thousands of years.

The reason to nevertheless work bottom up is because we know that that is the way matter has systematically built up in the early stages of our universe. Starting from the basic constituents that stepwise aggregate into complex structures on large scales turns out to be the true historical account after all. The universe cooled down in the course of its expansion. This means that thermal collisions between constituents became less and less violent, so that ever weaker and more subtle binding mechanisms could become effective in forming increasingly complex stable structures. These structures emerged as a result of the the four basic interactions and because the external conditions like temperature and density kept changing. Let us go through some of the very early stages guided by the events marked in Figure III.1.3.

The Planck and inflationary era. We discussed the very early stages of the universe in the section on Big Bang cosmology on page 66 of Chapter I.2. The true origin of our universe is hidden behind the curtain of quantum gravity

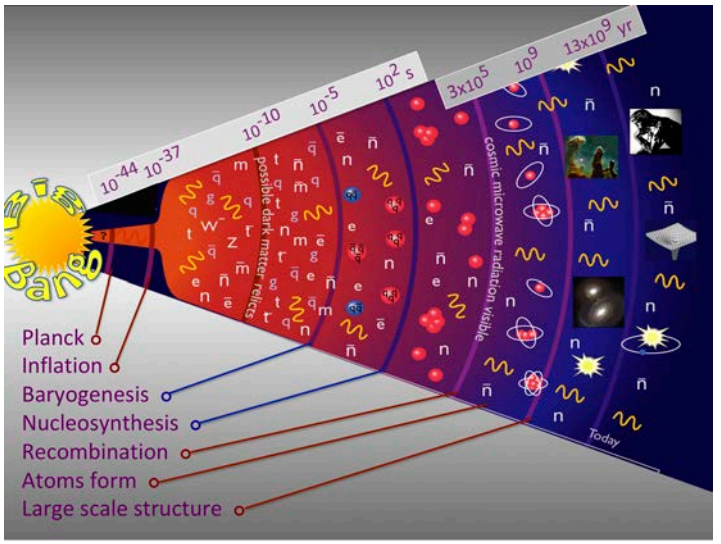


Figure III.1.3: *Cosmic evolution.* The figure shows the subsequent phases of the early universe, exhibiting matter organizing itself in ever more complex structures.

for which we do not have a satisfactory theory. That curtain obstructs our understanding of the universe for times smaller than the Planck time which is about $10^{-44}\ s$. So, what the Big Bang really is we don't know, but that such a dramatic event took place some 13.7 ± 0.2 billion years ago is beyond doubt. This was established unequivocally from observing the aftermath of it. A first grand event is the period of *cosmic inflation* where our universe scaled up exponentially thereby generating an enormous amount of vacuum energy and making it homogeneous, isotropic and flat. The picture is that the latent vacuum energy of the inflated universe was converted into all the (dark)matter and radiation that fill the universe today.

Primordial baryogenesis. Shortly after the Big Bang the universe was presumably filled with the most basic forms of energy: a *primordial soup*! Matter in the form of quarks, leptons, their antiparticles and many types of radiation. The strong interactions were operative; however, the quarks and gluons were not in a confining phase, but in the *quark-gluon plasma* phase we mentioned on page 195 in Chap-

ter I.4. A separate important question is the presence and role that dark matter may have played in the very early stages of the universe. This role strongly depends on what dark matter precisely is. What we know for sure is that it interacts very weakly with ordinary matter, and therefore it will not greatly affect the processes we will describe next. Ordinary matter and radiation are all interacting frequently enough to stay in equilibrium with each other. There is a simple rule, following on from special relativity that tells us that matter and anti-matter will *recombine*, and effectively annihilate each other if the temperature drops below twice the mass the particle type: $kT \leq 2mc^2$.

This in addition assumes that the density is large enough so that they will run into each other enough. Not much matter would be left if a slight asymmetry between matter over anti-matter did not develop at a very early stage, so that after the annihilation of all available anti-matter, a tiny surplus of matter (of 1 part in 10^9) remained and that is all the ordinary matter present in our early universe.

Primordial proton and neutron synthesis. When the universe was roughly 10^{-6} seconds old, the up and down quarks started binding into protons and neutrons due to the color force mediated by the gluon particles. The nucleon synthesis processes are

$$\begin{aligned} u + u + d &\rightarrow p \\ d + d + u &\rightarrow n \\ e + p &\leftrightarrow n + \nu \end{aligned} \qquad \text{(III.1.1)}$$

In this phase the universe was basically filled with a plasma consisting of protons, neutrons and electrons, and radiation consisting of photons and neutrinos.

Primordeal nucleosynthesis. After about 3 minutes the first nuclear fusion processes started to take place, the so-called *primordial nucleosynthesis* in which the lightest stable nuclei were produced like ^4He, ^3He and tiny amounts of lithium (^7Li) and beryllium (^7Be). The process stopped there, basically because there were no stable nuclei with a higher atomic number. The typical sequence of fusion steps 're:

$$\begin{aligned} p + n &\rightarrow {}^2D \ (Deuterium) \\ {}^2D + p &\rightarrow {}^3He \\ {}^2D + n &\rightarrow {}^3T \ (Tritium) \\ {}^2D + {}^2D &\rightarrow {}^4He \\ {}^3T + {}^4He &\rightarrow {}^7Li \\ {}^4He + {}^3He &\rightarrow {}^7Be \\ &\dots \end{aligned} \qquad \text{(III.1.2)}$$

Note that the process proceeded via unstable intermediates such as the hydrogen isotopes, deuterium and tritium, mostly ending up in stable ^4He nuclei. After the first fifteen minutes the cosmic abundances settled to about 75% hydrogen ($H = p$) and 24% helium-4. The prediction of these primordial cosmic abundances was one of the important successes of using quantum (nuclear) theory in the context of the early universe. Many others were to follow.

Gravities opportunity: the seeds of large-scale structure. Only after about $300,000$ years the simplest atoms would form, meaning that the electrons would combine with the aforementioned nuclei to form electrically neutral atoms. At that point the universe was filled with a gas of neutral atoms. The photons decoupled, and the gravitational force became dominant. Inhomogeneities corresponding to local maxima in the mass density of particles attracted other particles more strongly than the low density regions and therefore high density regions started to build up mass. From a gravitational point of view all masses attract each other, and the more mass the stronger the attractive force. This means that pockets where the energy density is more than average will grow. These early density inhomogeneities are the seeds of the large-scale structure in the universe.

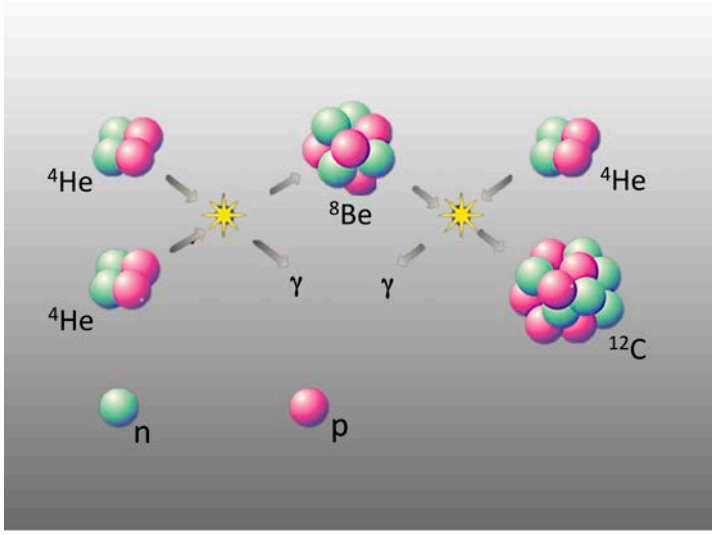


Figure III.1.4: *Carbon production.* It is shown how carbon nuclei were produced in the universe by successive fusion processes of $^4\mathrm{He}$ inside stars.

From stardust we are made. In the center of these ever-denser clouds, pressure and temperature started building up locally reaching again high temperatures of millions of degrees. This gave rise to a next round of nuclear fusion processes. That is how slowly the diverse array of chemical elements in nature was created in the core of many generations of stars, and the stockpile of basic chemical elements, indispensable for the later chemistry of life, was built. The truth is that all of us are made of stardust! It is interesting to be aware of the fact that this process took billions of years because several generations of stars were needed to build the heavy nuclei. And the fact that our expanding universe has to be old explains why it is also big and cold. It *has* to be, otherwise we could not be there to observe it. What feels like an utter inhospitable environment turns out to be necessary for life to be possible in the first place.

We see from the periodic table that in principle by adding on $^4\mathrm{He}$ nuclei, elements like beryllium and the all-important carbon and oxygen can be reached, as indicated in Figure III.1.4. For example:

$$\begin{aligned} {}^4\mathrm{He} + {}^4\mathrm{He} &\rightarrow {}^8\mathrm{Be} + \gamma \\ {}^4\mathrm{He} + {}^4\mathrm{He} + {}^4\mathrm{He} &\rightarrow {}^{12}\mathrm{Carbon} + \gamma \\ {}^{12}\mathrm{Carbon} + {}^4\mathrm{He} &\rightarrow {}^{16}\mathrm{Oxygen} + \gamma \end{aligned} \quad \text{(III.1.3)}$$

The way carbon is synthesized is remarkable to say the least. The effectiveness of the processes above is due to a subtle resonance which amplifies the second process. It remains mysterious that on the one hand all of life is carbon based, whereas the actual production of the carbon itself was a process depending on a delicate balance of values of the constants of nature. From this point of view, one is tempted to conclude that life is a miraculous coincidence!

Chemical element	Milky way	Solar system	Earth crust	Human
H	73.90	70.57	0.14	10
He	24.00	27.52	-	-
O	1.04	0. 59	46.00	65
C	0.46	0.30	0.03	18
Ne	0.13	0.15	-	-
Fe	0.11	0.12	5.0	6×10^{-4}

Table III.1.1: *Mass abundances.* Abundances (in %) of some common chemical elements at different extraterrestrial and terrestrial levels.

In Table III.1.1 you see what happened to the original galactic abundances, like in our Milky Way, on their way to become tiny parts of our physical bodies. The explanation of how these changes came about goes beyond the scope of this book.

Molecular binding

Atoms are electrically neutral because the positive charge of the nucleus is exactly cancelled by the negative charge of the electrons. Yet the charges are not exactly on top of each other so what you find if you go to short distances is that there are residual electromagnetic interactions (like dipolar forces) that become dominant. These residual interactions are to a large extent responsible for the fact that atoms bind in such a rich diversity of structures, be it molecules of varying complexity, or solids, or other types of condensed states of matter.

Repulsion versus attraction. Interactions are the mother of binding and binding is the father of structure. The secret of building spatially extended structures resides in the fact that the binding between atoms or molecules is the outcome of a delicate balance between a repulsive force that dominates at small distances and an attractive force that dominates at large distances. The typical behavior for the energy U of a pair of atoms as a function of their separation r is given in Figure III.1.5. Understanding the curve is not hard. Imagine releasing a marble on the energy curve, then starting at a small r it would roll away to large distances (that is the repulsive part of the interaction), but starting for large r it would role towards the origin (the attractive part). So, if the particle were to experience some friction then irrespective of where you start the marble would always end at a separation $r = r_0$, where the potential energy is minimal. This picture reminds us of the atomic binding of Figure I.4.5 at least in a qualitative sense. We conclude that also in this domain stability is based on a compromise between attraction and repulsion. This is a feature underlying the formation of structure on most levels of complexity.

Van der Waals binding. The basic attractive interatomic force is the Van der Waals force after the Dutch 1910 Nobel laureate Johannes Diderik van der Waals. It even works

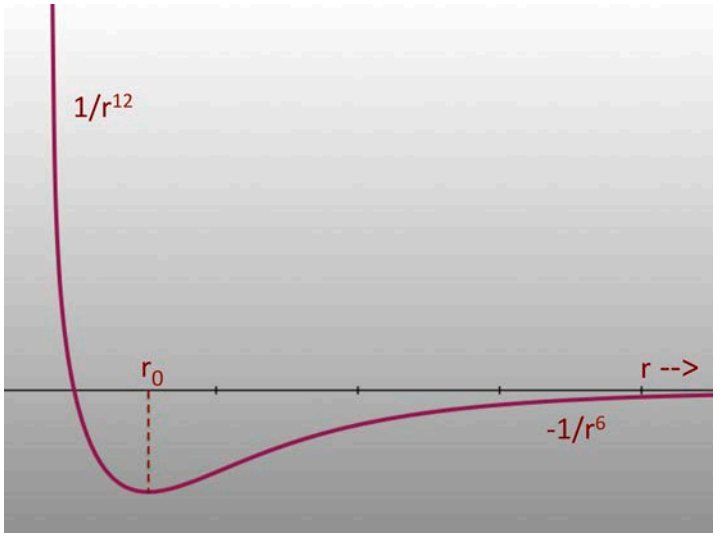


Figure III.1.5: *The interatomic interaction potential.* The interaction potential of two hydrogen atoms as function of their distance. For short distances the force is repulsive but for long distances attractive. This behavior is a consequence of the sharing of electrons which implies that a negative charge cloud forms between the two positively charged nuclei. The minimal energy configuration is achieved for a distance r_0. So free hydrogen spontaneously forms a gas of diatomic molecules H_2.

between two atoms that are called 'inert' like argon or neon. They have completely filled shells which means the charge cloud is spherical. However, if they get close these clouds become deformed and the molecule develops an (induced) dipole moment which just means that the resulting plus and minus charges have different spatial distributions. The induced dipole moments lead to a weak attractive force between the atoms. It is weak because the interaction potential drops off as $\sim 1/r^6$ that is much faster thus than the Coulomb potential ($\sim 1/r$) between two opposite charges. On the other hand, if the atoms are attracted they cannot come too close because then the electron clouds start overlapping and that causes a strong repulsion and a steep rise of the potential for short distances ($\sim 1/r^{12}$). That repulsion is due to the Pauli principle which holds for the electrons: it provides a hard core for the interactions. This

potential is depicted in Figure III.1.5. At low temperatures the Van der Waals interactions may lead to the formation of a solid where all the atoms form a regular array, and the nuclei occupy the sites of a crystal lattice.

Polar (or ion) binding. Atoms have a certain number of electrons which form a charge cloud around the nucleus. The electrons subsequently have to occupy different states that is why the charge clouds differ from atom to atom. Now for the chemistry of atoms for example which molecules they can form, the shape of the clouds is all-important. The number of *valence electrons* is the number of electrons in the highest unfilled shell. The tendency of atoms is that they like to fill their outer shell. They can do that basically in two ways: one is that they can pick up the electrons of another atom in which case the atom that gives away electrons becomes a positive ion and the one that takes extra electrons becomes a negative ion. The ions have the same old nucleus but have a net charge because of an electron surplus or deficit. Clearly the ions made through this 'social' mechanism of giving and taking have opposite charges and will be attracted to each other because of the Coulomb force between them. But again, at small distances the repulsive interaction of the clouds takes over, and qualitative features of the picture of Figure III.1.5 remain valid.

A lot can be said based on the location of the atoms in the periodic table in particular the column they are in. Take the elements in the first column like hydrogen for example, they have one electron in the outer shell. As it happens these atoms are actually quite social: they are willing to give away their electron and to turn into a positively charged ion. Complementary behavior is obserbed in certain elements in the one but last column, like chloride (Cl), that like to receive an extra electron to fill their outer shell and turn into a negative ion. So indeed, we see *polar binding* between atoms in the first column and the one-but-last column. And we see many well-known elementary molecules like HCl (hydrochloric acid) and NaCl (kitchen salt) that are held together this way.

Covalent binding. Simple atoms like hydrogen, oxygen or nitrogen, which are the main components of ordinary air, are bound in pairs. The question is how the pair-binding in the diatomic gases precisely comes about. How can it work because there are no ions to be formed? In these cases a different mechanism is operative that is also quite 'social', as it is based on the notion of *sharing*. Once close enough, atoms can lower their energy by sharing outer electrons; they spread as it were their negative charge clouds over the two nuclei, by sharing electrons. The cloud is mostly concentrated between the nuclei and that means that these become attracted to the cloud and therefore to each other. The binding that results from this mechanism is called *covalent binding*.

We have mentioned that what matters are the shapes of the charge clouds corresponding the outer (or valence) electron orbitals. They tell us a lot about the geometrical patterns of molecules and materials. On the other hand, once we realize that the atoms are composites of nuclei and electrons and therefore by themselves complex objects, we should not be surprised to learn that in the behavioral diversity they exhibit, much will depend on the details of the atoms in question.

Hydrogen bonds. Once you know how atoms form molecules there is the next step up, which is to understand how molecules bind with each other or in case they become large, how they interact with themselves to produce more and more elaborate molecular structures. Here one exploits more intricate mechanisms that will do the job. A well-known example of this is the so-called *hydrogen bond* that plays a vital role in organic chemistry and therefore also biochemistry. It is based on the idea that molecules, or parts of molecules, may also behave like electric dipoles and therefore lead to an attractive force. The term hydrogen here refers to the fact that hydrogen, when it binds to a strong electronegative atom such as oxygen or nitro-

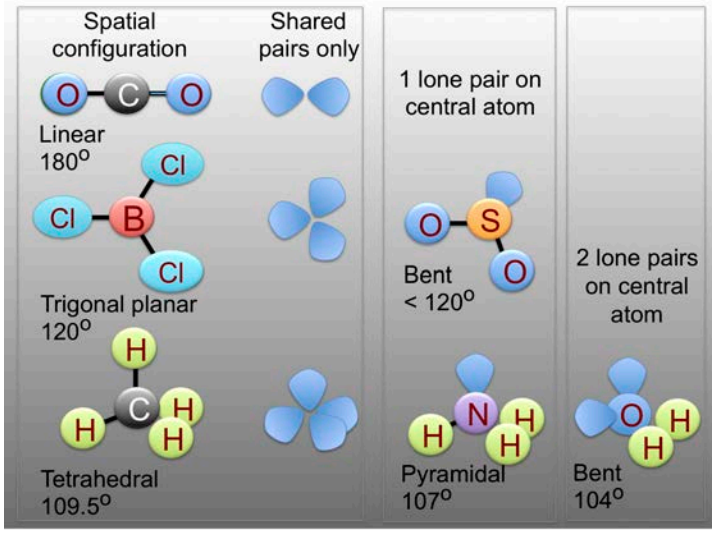


Figure III.1.6: *Molecular shapes.* We have depicted the spatial geometry of the atoms forming a molecule, and the charge clouds corresponding to the shared and lone electron pairs.

gen, like in water, gives a polar molecule that binds through these hydrogen bonds. This type of binding is what keeps the water molecules together in the liquid, and it for example explains the relatively high boiling temperature of water. The hydrogen bond is thus structurally similar to the Van der Waals force, but it is stronger. These bonds play a vital role in understanding the spatial geometry of complex biomolecules.

It's all quantum plus electrodynamics. All this being said, I like to stress that all chemical binding mechanisms are a product of two fundamental ingredients. One is the set of underlying quantum principles as expressed by the Schrödinger equation, and the other set is formed by the laws of electromagnetism governing the forces between charges. It means that if – as is often done in practice – we were to put the constituents and their basic electromagnetic interactions in the Schrödinger equation and let a powerful computer turn the crank we would generate the structures we observe. Such calculations show that the theory is correct and have great value for applications. They do however not replace or satisfy our need to understand the basic physical and chemical mechanisms. Scientists have introduced many so-called forces and effective interactions and bonds, exactly because they provide a kind of elementary toolkit to effectively explain and predict chemical behavior. But we should remember that all of those new forces are nothing but residual electromagnetic interactions between objects like atoms or molecules or chemical 'groups' that have intricate charge distributions determined by the laws of quantum theory. It's all a matter of shapes and these shapes can be described as 'multipolar fields' of which the dipole is the simplest example. The quantum laws are strong, accurate and universal, and even though they don't allow us to understand all of chemistry directly from first principles, they do allow us to comprehend in detail the basic mechanisms that in a subtle balance give rise to the elaborate chemical structures we observe in nature.

The miraculous manifestations of carbon

The plug and play of organic chemistry. In this subsection we take a closer look at the element carbon and the remarkable structures it can form all by itself, as displayed in Figures III.1.7 and III.1.8. We start simple and add more complexity along the way.

The spatial geometry of simple molecules. Because the carbon atom sits in the fourth column of the periodic table, it has four valence electrons to share. Hydrogen has one to share so carbon can bind to four hydrogen atoms to form a methane CH_4 molecule, which as you probably know is a strong greenhouse gas molecule. Both atoms are happy because they made a perfect match in the one to four ratio. What about the other bad guy, carbon dioxide CO_2? Well, now the carbon shares two electron pairs with each of the oxygens to optimize its sharing strategy. And what about H_2O, just innocent water? Well, the oxy-

gen clearly shares one pair with each of the hydrogens and there are four non-paired electrons left on the oxygen.

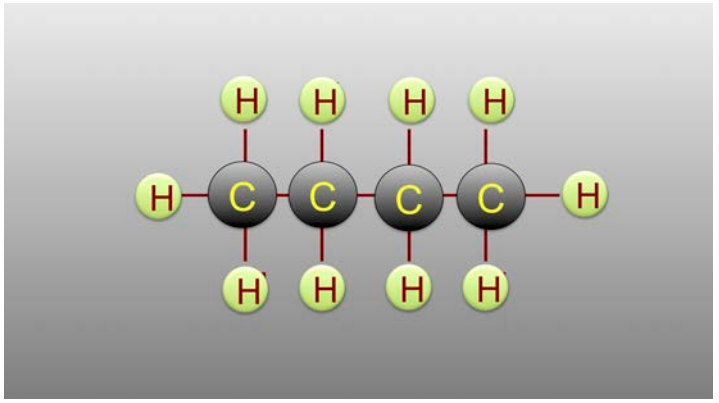
The next question that naturally arises is what do these molecules look like? Can we from the binding mechanism decide what the spatial configuration will be? For simple molecules this is indeed the case as is shown in Figure III.1.6. The resulting shape follows from the mutual repulsion of the negatively charged electron clouds, which try to avoid each other as much as possible.

Shapes of simple molecules. So, for the methane or CH_4 it should not come as a surprise that it forms a perfect tetrahedron with the carbon nucleus at the center and the hydrogen nuclei at the four corners. The clouds on the bonds indeed maximally avoid each other meaning that the bonds will make angles of 120 degrees. For CO_2 there are two double bonds and we expect a linear structure with the carbon nucleus in the middle right in between the two oxygens. A detail is that indeed a double bond defines a plane, The two double bonds mutually repel and therefore the plane connecting to the first oxygen will be perpendicular to that connecting to the second. And what about the water molecule H_2O, is it also linear? Here there is another ingredient: the four leftover electrons of the oxygen form a cloud also attached to the oxygen. So, in fact there are three clouds that will lie in a plane, and as the clouds are not identical the H_2O molecule has a bent structure. The lone pairs tend to be bulkier and therefore push the peripheral atoms down so that the angle between them will to be smaller than in the symmetric case. That explains why the two bonds to hydrogen make an angle, not of 120, but of about 104 degrees.

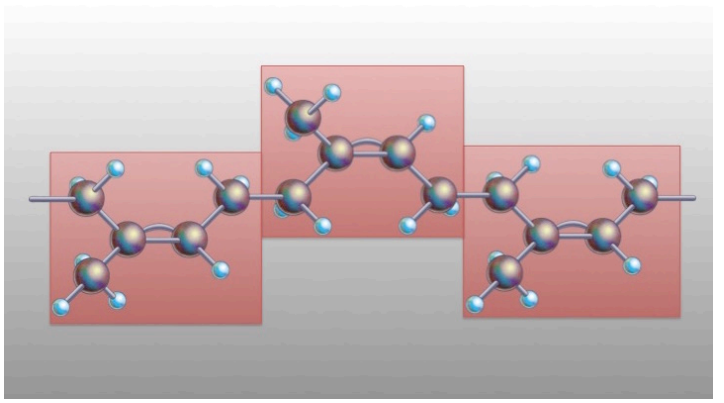
Greenhouse gases. Carbon dioxide is made by burning carbon containing materials. It is an enormously useful chemical compound but the problem is that we have produced and still produce far too much of it. It plays a hazardous role in our atmosphere as it is a greenhouse gas. This is the case, because molecules which have a certain structural complexity (like carbon dioxide, methane, but also water vapor) have many low energy, oscillatory quantum mechanical modes in which they can absorb and (re)emit radiation. In particular, modes corresponding to heat radiation. So, the heat that is coming from the Earth's surface after being absorbed from the sun, or heat produced by human activities, gets absorbed by the CO_2 blanket in the atmosphere, and then reemitted. But the reemission is isotropic, meaning the same in all directions, and therefore half of the reemitted heat goes back to the earth and that is why the earth heats up.

Photosynthesis. One way to get rid of CO_2 is through vegetation; plants absorb carbon dioxide from the air, and in a process called *photosynthesis* combine it with water and light (photons) from the sun to produce carbohydrates and the oxygen we need in a process which can be summarized as $CO_2 + H_2O \rightarrow [CH_2O] + O_2$. Water vapor in the air certainly does affect the greenhouse effect in that it increases the warming up caused by carbon dioxide considerable. However, water is engaged in all kinds of other climatological cycles like cloud formation and rain that make its role essentially different, the vapor concentration in the atmosphere changes by large amounts on a short scale of days or weeks.

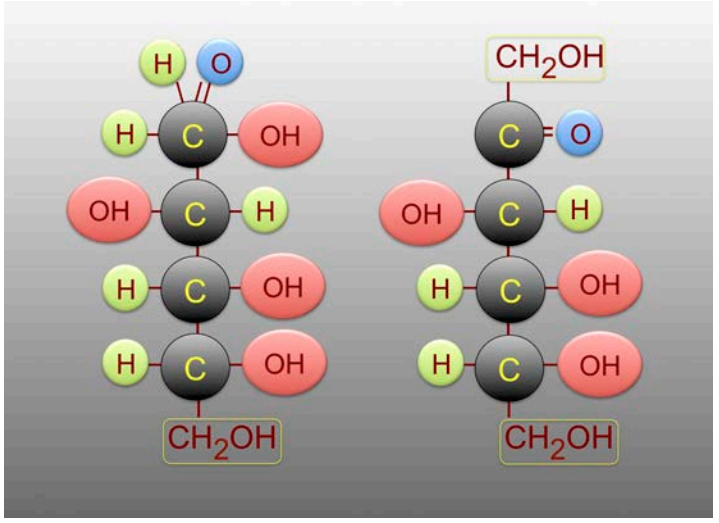
Carbohydrates. Once you realize that carbon has four binding sites available you realize that there are extremely diverse ways to combine these molecules Carbon is an ideal example of a basic building block. And nature learned to play with it. Imagine you start with a tetrahedral *methane* CH_4 molecule, and you replace one hydrogen by another carbon then that is also a compatible configuration. Continuing this process two more steps you get the *butane* molecule of Figure III.1.7(a). It is evident that *carbohydrates* like C_kH_{k+2} actually can in principle form for any value of k. These molecules correspond to long linear chains.



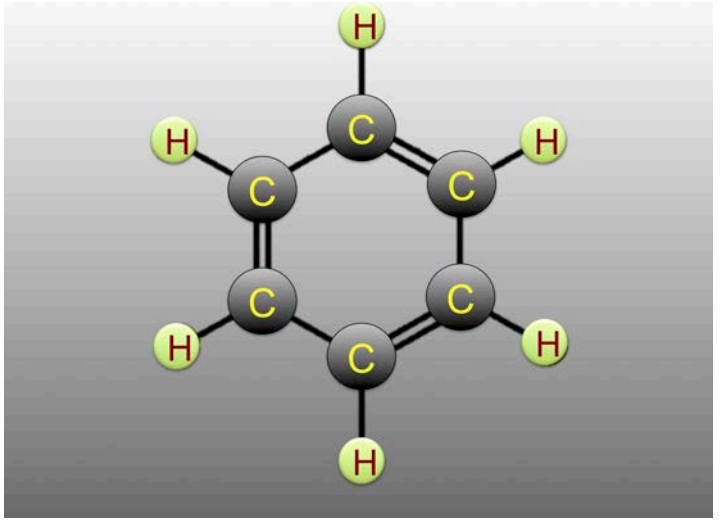
(a) Carbon has the powerful property that it can form long linear chains with hydrogen atoms on the side. This is the highly flammable gas butane for example.



(b) A polymer is a linear chain made up of identical units.



(c) The common sugars or carbohydrates glucose (l) and fructose (r). These have a chirality or handedness; there are two forms. The case where the bottom group is on the left or the right, is like a left or right shoe. They form mirror images that cannot be rotated into each other.



(d) If you can make chains you also can make cycles without extra ingredients. This is the benzene molecule C_6H_6 featuring the famous hexagonal ring structure with three double and three single bonds.

Figure III.1.7: *Miraculous carbon.* Carbon plays a central role throughout organic chemistry. With its four bonds it is remarkably versatile and can make linear, planar or 3-dimensional structures.

Polymers. One can go one step further and build long linear molecules that are repetitive. Such long chains of identical or similar units are called polymers as shown in Figure III.1.7(b), and it is a world on its own, to design polymers in such a way that they exhibit dedicated chemical properties, with particular applications in mind. This is what a substantial part of the bulk chemical industry is about.

Ring structures. There is not only the possibility of open carbon chains, but you can also imagine the formation of cycles or closed chains like the so-called *benzene ring* C_6 which nature discovered and used over and over again. Ring structures like *cyclopentane, cyclohexane* and their polygon shaped relatives play an important role in the biochemistry of the base pairs in DNA and also in the *amino acids* from which the *proteins* are built. Furthermore, they are 'bread and butter' for the chemical and food industries.

Nano physics

Nano science. Carbon composites don't stop in the one-dimensional world of chains and cycles. Nothing keeps it from engaging in three valent bindings, meaning that a C atom has not just two C neighbors, but three that form an equilateral triangle. Such a connection opens the possibility of making two-dimensional structures with the topology of planes, tubes and balls, and two-dimensional surfaces that have holes in them, the simplest one being the torus or donut.

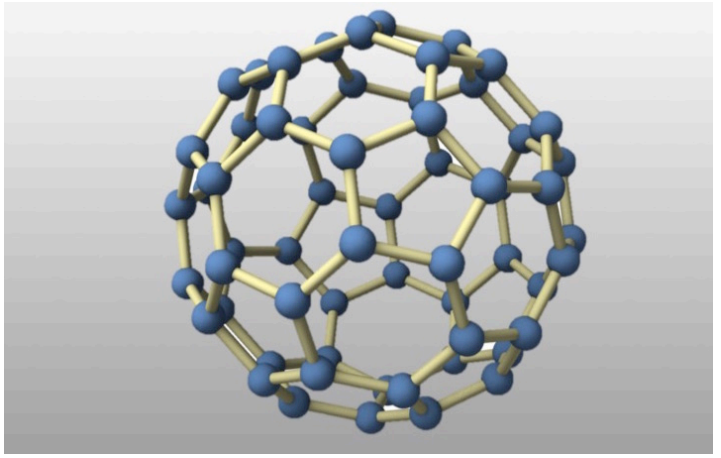
Mesoscopics. With the carbon structures we just mentioned we enter the unfolding world of nano-science and technology, where one is dealing with molecular structures on a nano scale, so typically involving up to a few hundred atoms. This domain is also called *mesoscopic*, just in between the macroscopic and microcosmic worlds.

Nature's LEGO. Every parent remembers the thrill of what happens after you hand a group of playful children a big box of the most basic LEGO pieces. It is amazing what kind of stable and metastable structures they come up with. In this sense evolution is like a room full of children with an overdose of LEGO pieces, and once you realize that, those elaborate carbon structures become little more than the inevitable outcome of a childlike but powerful methodology called trial and error.

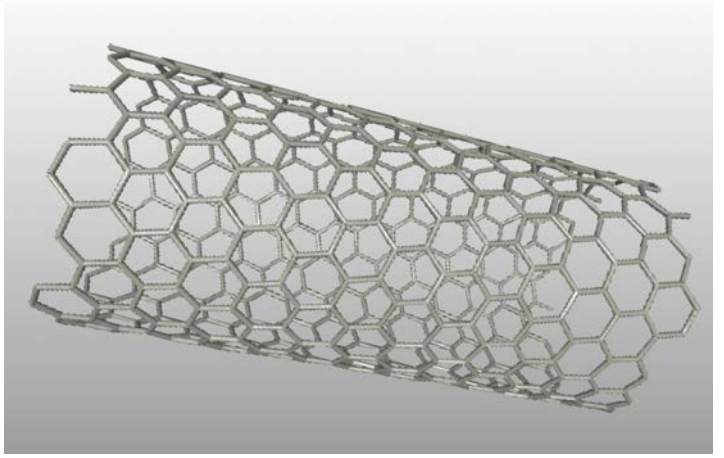
Buckyballs. A most remarkable discovery was the buckyball or C_{60} gigantic molecule that is spherical rather than linear and made up of alternating pentagons and hexagons (see Figure III.1.8(a)). It was predicted by theoretical calculations to be extremely stable. Such large carbon molecules (not only C_{60} but actually a whole range going from C_{40} to maybe C_{240}) are now called *fullerene*s. This name refers to Buckminster Fuller, the American architect who pioneered the design and constructions of geodesic domes.

Nano tubes. Closely related are the nano-tubes depicted in Figure III.1.8(b) which have attracted a massive amount of attention because of their many potential applications. These tubes are thin: the smallest have a diameter of only a few nanometers. This makes them extremely strong in proportion to their weight. Large nano-tubes are hard to make and this has so far hampered their large-scale application in technology. Let us finally mention the materials that are only made from carbon atoms.

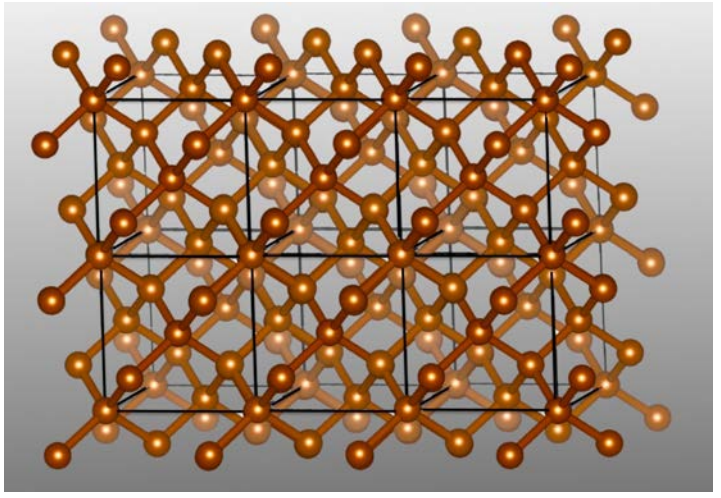
Diamond and graphite. As each C atom has four C neighbors, naturally located at the corners of a tetrahedron, it allows for the formation of wonderful three-dimensional lattices. One of those is quite exquisite indeed, because it is the diamond lattice. Diamond is pure carbon in a splendid guise, as it is extremely hard, highly transparent and very expensive. Diamond has relatively high density $(3.5\ \mathrm{g/cm^3})$, does not conduct heat or electricity and is insoluble in any solvent.



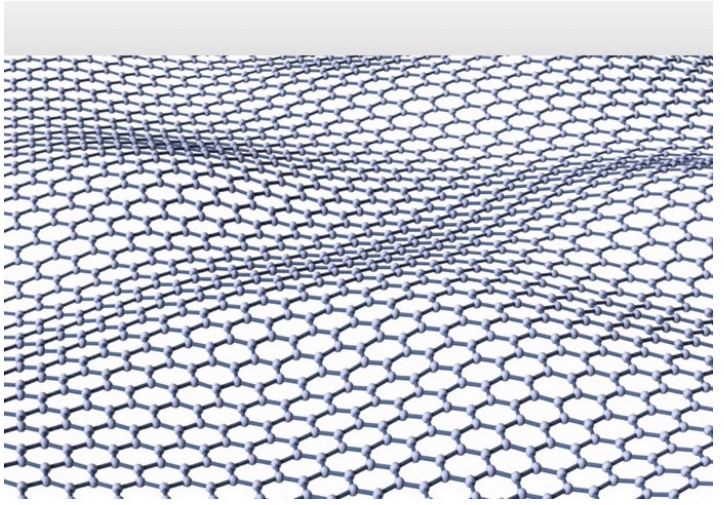
(a) The football shaped C_{60} molecule is an example of a *fullerene* after Richard Buckminster Fuller, the architect and pioneer in designing and building of geodesic domes.



(b) A carbon nanotube.



(c) The structure of the covalent diamond lattice made with carbon atoms on all sites.



(d) Amazing graphene: only one molecule thick, and yet the strongest planar material. It is also transparent and an excellent conductor.

Figure III.1.8: *Carbon structures.* Some of the miraculous manifestations of carbon that all manifestly exploit the hexagon as basic building block.

Are there other three-dimensional carbon structures possible? Yes, there is one, much more common than diamond, and that is *graphite*, the stuff that sits in your pencil and makes drawing so easy because it is totally opaque (black), soft and cheap as well. These properties follow from the fact that graphite forms easily, it corresponds to a stack of two-dimensional honeycomb planes that are relatively weakly bound. Graphite is soft and greasy, it is relatively light ($2.5\ \mathrm{g/cm^3}$), a good conductor of heat end electricity and is soluble in most solvents. How different can members of one family be!

Graphene. Let us finally mention the recently discovered miraculous material called *graphene*; this is a perfect two-dimensional hexagonal honeycomb sheet which turns out to be extremely strong in spite of being only a single atomic layer (see Figure III.1.8(d)). It is furthermore transparent and has high thermal and electric conductivity. This highly unusual combination of qualities singles this material out for many exceptional applications in the future, varying from wearable electronics and displays to fancy wrapping materials. It may strike you that the structure is just like a single layer of graphite. The story goes that the Russian physicist Andre Geim and his student Konstantin Novoselov who received the Nobel prize for their groundbreaking work on graphene in 2010 made the first specimen just drawing with a pencil on the sticky side of sellotape.

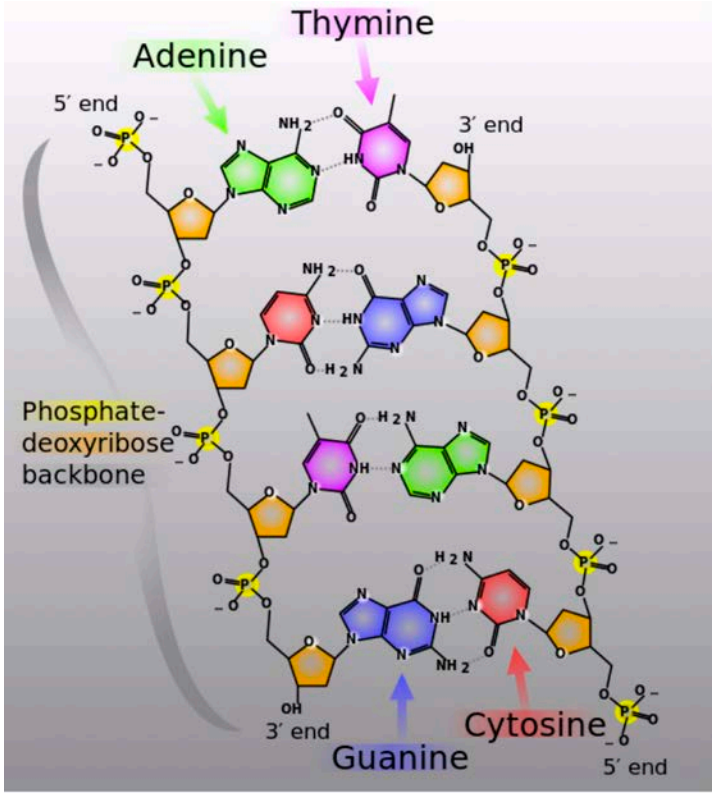


Figure III.1.9: *The chemical composition of DNA.* A fragment of the double-stranded DNA molecule. The picture also gives the molecular structure of the base molecules with the four-letter code assigned to them. The four letters $\mathrm{A}, \mathrm{T}, \mathrm{G}, \mathrm{C}$ are strictly paired as $\mathrm{A}-\mathrm{T}$ and $\mathrm{G}-\mathrm{C}$. The pairs are relatively weakly bound by hydrogen bonds indicated by the dotted lines. The DNA of the human genome contains about 3 billion base pairs, which contain among other things the genes that encode about 20,000 proteins. (Source: Wikipedia)

The molecules of life

The pinnacles of molecular structure are the molecules of life such as nucleic acids and proteins. It seems somewhat far-fetched to present these in an elementary book on quantum theory. The reason I do is that the structural hierarchy, as far as single molecules are concerned, really ends right there. And these structures are basically dictated by quantum theory. Therefore, including them gives our review of the molecular hierarchy a sense of completeness. Let us therefore briefly summarize some structural aspects and not talk about the functional part. As a matter of fact the real tasks in the living cell are mostly performed by complex networks of proteins, and that is a level of emergence that transcends the one fully fixed by the basic laws of physics.

The complexity of biomolecules is relative in the sense that again it is a structural level in which a limited number of

particular building blocks are used over and over again. Nature is brilliant in figuring out ingenious ways to apply a given structural element in many different ways. The structure of biomolecules is modular and the huge diversity is not as much in the variety of constituents, as it is in the way they are put together on a modular level.

The DNA molecule. A well-known example is the DNA molecule which is made of tens of billions of atoms. But its structure is highly repetitive so that one only has to show a little piece to see and understand what the building principles are. And once the architecture of the molecule is understood it is not so hard to explain the way it functions either. The structure of the molecule was discovered in 1953 by Francis Crick, James D. Watson at Cambridge University and Rosalind Franklin at King's College London. The Nobel prize for Physiology or Medicine was in 1962 awarded to the first two and Maurice Wilkins, a collaborator of Rosalind Franklin in London.

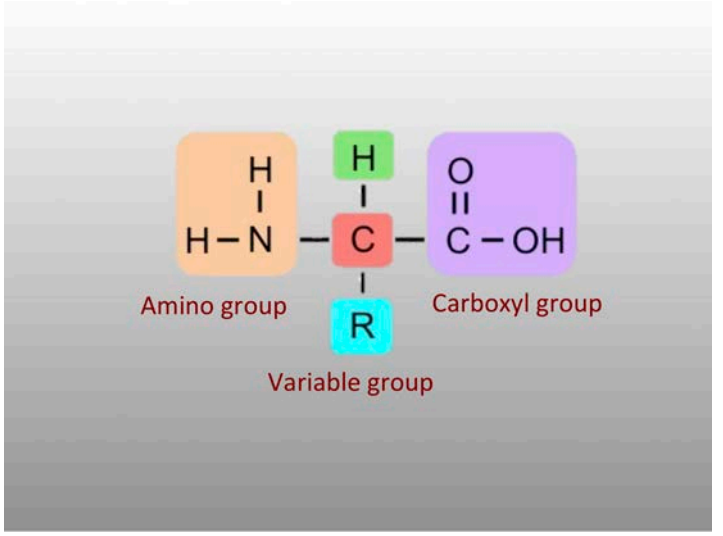


Figure III.1.10: *Amino acids.* The generic structure of an amino acid, with its amino and carboxyl groups. In the center is a specific group that characterizes the particular amino acid. Proteins are basically linear chains of amino acids.

We have illustrated a small segment of the molecule in Figure III.1.9, and it is clear that the molecule features two long strands that are kept together with hydrogen bonds to make a sort of ladder. The stiles of the ladder are just a backbone of some sugar that repeats itself some three billion times. The rungs of the ladder are made of pairs of *nucleobases*, of which there are only four, called *adenine* (A), *thymine* (T), *guanine* (G) and *cytosine* (C). It is the order in which these four types of rungs appear in the ladder which encodes the heritable traits of living organisms. There is a strict pairing namely A always comes with T and G always with C, so if you know the left half of DNA it is easy to construct the complementary right half of the molecule. And it is this deterministic feature that allows us to understand how the heritable information can be reproduced after the cell division where the DNA molecule splits and the left and right half move to the two different daughter cells, which then are completed by synthesizing the complementary half within the daughter cell. The chemistry is in fact rather simple but extremely effective. If you think of the genetic information stored in DNA as a piece of text written in a four-letter alphabet of some 3 billion letters long, then that would maximally amount to $\mathrm{N} = 4^{3\,000\,000\,000}$ possibilities, which corresponds to six billion ($= {}^{2}\log\,\mathrm{N}$) bits of information. That amount of data would easily fit on a DVD or USB stick, in fact a good deal less because most of the information is highly repetitive and not conserved at all and therefore believed not to be that important. Yet as we are talking about important hereditary data, we should realize that the same DVD is sitting in every nucleus of every cell of our body – you should imagine that you are carrying around trillions of backups of your genome. I must admit that it makes me feel some kind of important, The DVD of my personal 'feel good' movie is not for sale but nevertheless made in huge quantities. This is how the discovery of a deep secret of life ended up being a little more than a paean to painstaking reductionism.

Translation of DNA information to protein structure. DNA is crucial for the organism but it doesn't do very much,

from a chemical point of view, it is not very active. It functions as a template from which the data corresponding to a *gene* are transcribed by RNA molecules that also carry it outside the nucleus of the cell where the instructions are then performed by *ribosomes* (some enzyme) to translate the four-letter code sequence of the genes as a sequence of three-letter *codons*. A codon encodes for a specific amino acid and the codons therefore form the *genetic code* The ribosomes produce from that sequence of codons a linear chain of *amino acids* corresponding to a specific *protein*. This process is schematically represented in Figure III.1.11. The number of different amino acids that can be encoded by a three-letter codon (word) with the four-letter alphabet, can never be larger than $4^3 = 64$. In fact, there are only twenty-one of them but most of them are represented by several different codons. This redundancy makes protein synthesis more fault tolerant against copying errors.

To make the structural hierarchy explicit and complete, I have displayed the generic structure of the amino acids in Figure III.1.10. Because of their modular structure they are in fact quite similar, consisting of an amino and carboxyl group and a specific variable group in the center. This group may contain five and six cycles and combinations thereof, somewhat similar to what we saw in the DNA segment of Figure III.1.9. A protein is just a linear sequence of amino acids that may run from ten to hundreds for small genes to hundreds of thousands for the big ones. And because of their characteristic charge distributions these proteins start to fold up in all kinds of interesting ways, as schematically indicated in Figure III.1.12. This is called the secondary structure, where one distinguishes so-called α *helices* and β *sheets* and simpler strings in between such as *turns* or *coils*. The helices are curled up and the sheets are more planar again with two strands bound by hydrogen bonds. The helices and sheets making up the protein are then again folded in characteristic ways into complicated and beautiful three-dimensional geometrical structures (see the rather random selection in Figure III.1.13). And again it turns out that their shapes determine to a large extent what biological functions the protein can perform.

Curling up. We should be aware of the fact that the gargantuan DNA molecule, which has a typical length say of 3 billion times a few nanometers ($= 10^{-9}$ m) equals some meters, apparently fits in a cell nucleus with a typical size of 10 micrometers ($= 10^{-5}$ m). This fact implies that nature must have developed some very clever folding tricks to make this possible. This is a generic feature of the big molecules of life, they are folded up in smart and elegant ways, and the way they are, usually tells us a lot about the biological function they may perform. DNA for example is curled up in different levels, first in small curls, then the curled up molecule curls up once more and then again etc... Like what certain phone cords do when you don't want them to. But to read the code corresponding to a gene, the corresponding part of the DNA molecule must be made accessible, i.e. certain genes have to be 'turned on', depending on what is needed in that particular cell at that time and place.

Epigenetics. At this point we enter the domain of *epigenetics* where one tries to understand how the gene expression in the organism is exactly regulated by means of other chemical mechanisms using *histones* and *methylation*. There are indications that also the methylation of the DNA is conserved, which means that it is somehow encoded in the DNA. It has been suggested to add a fifth letter to mark its positions along the molecule. Unsurprisingly, several meta-levels of regulation are operative to get from the *genotype* of the organism to the *phenotype*, to get from our DNA to who we are as an integrated being. Whether the development of an organism is primarily nature or nurture, chemistry is the language in which the explanation will ultimately be cast.

Conclusion. In this chapter we have shown how the complex hierarchy of matter came into being during the early

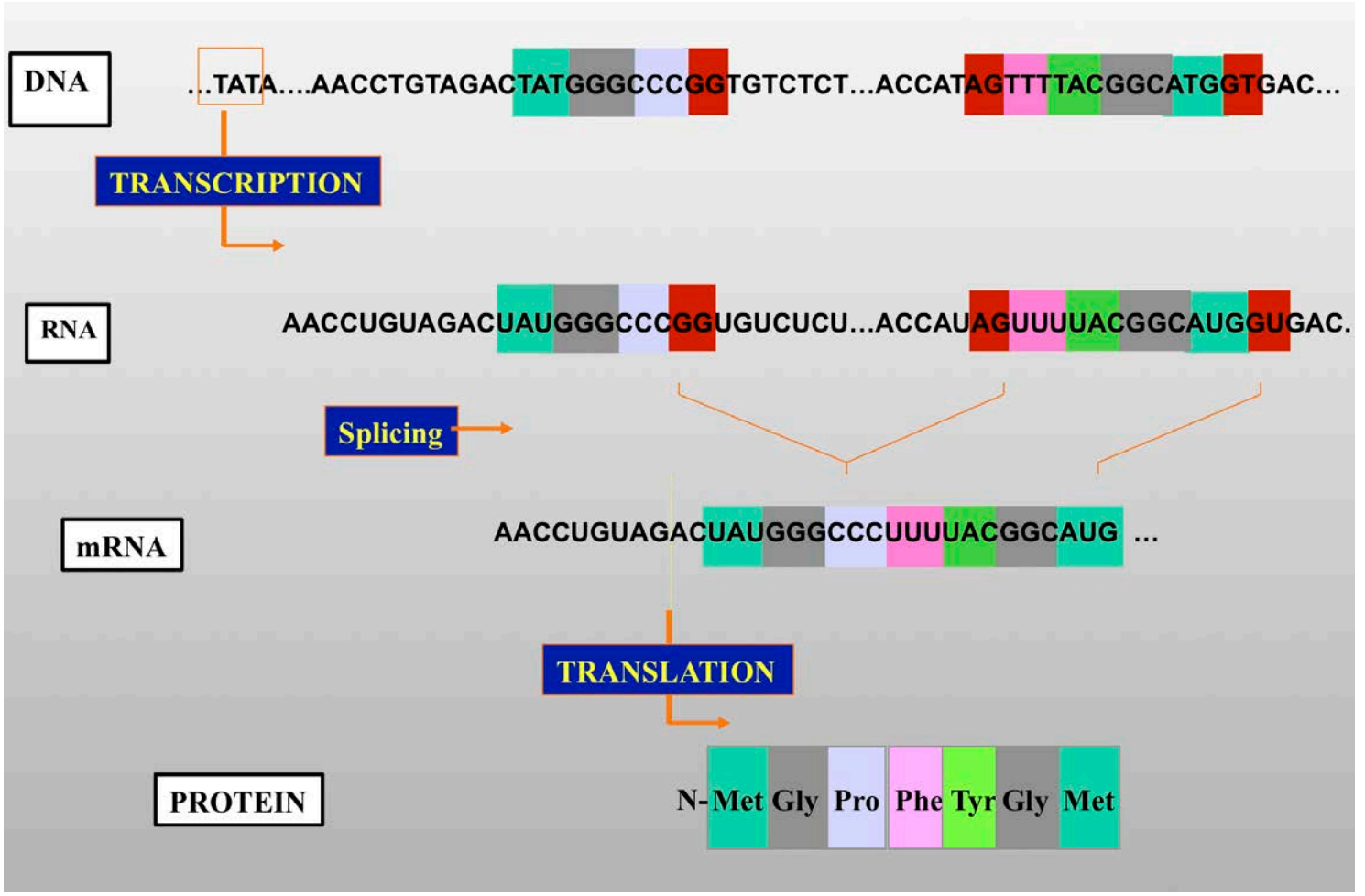
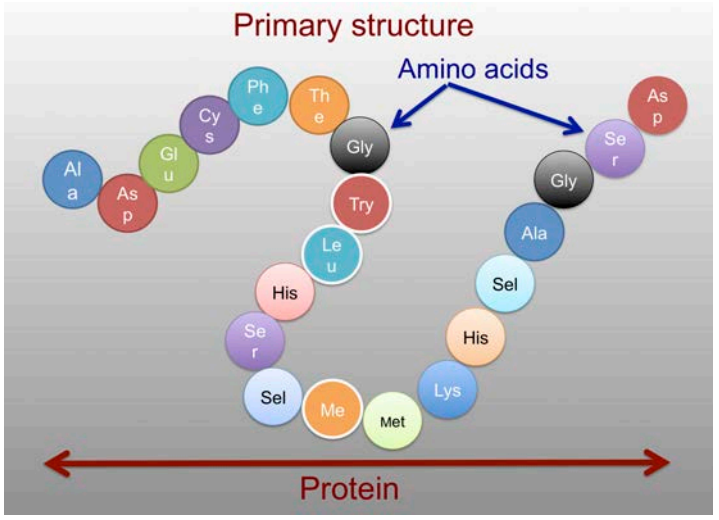
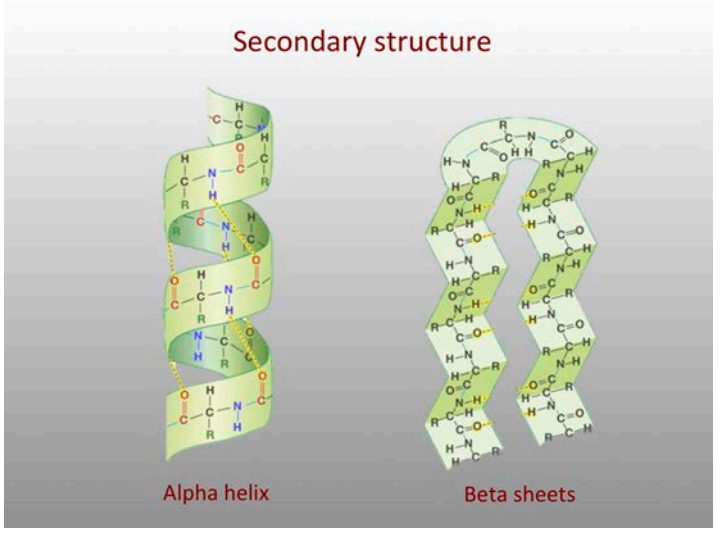


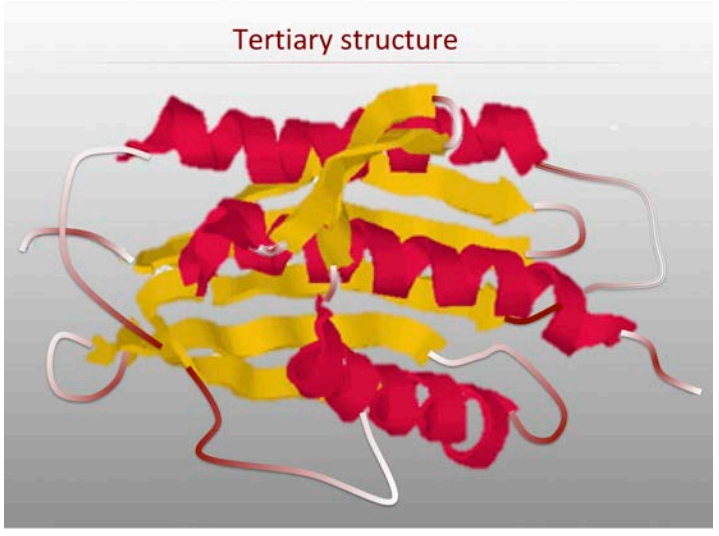
Figure III.1.11: *From DNA to proteins.* A schematic of how the linear four-letter code of DNA strand gets translated into a linear sequence of amino acids that form a protein. The four-letter code is copied on a single strand RNA. After splicing, which means cutting and copying the various pieces of the gene to a single sequence on a messenger RNA molecule, the messenger goes outside the nucleus of the cell. There the letter sequence is translated by Ribosome enzymes and the protein is synthesized. Each subsequent three letter sequence (called a *codon*) from the RNA gets translated into one of twenty-one amino acids, see Figure III.1.10.



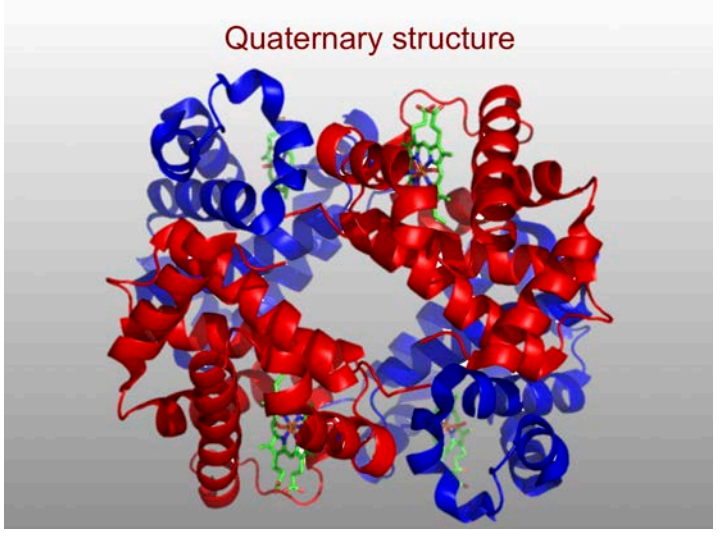
(a) Primary structure as a linear chain of amino acids



(b) Secondary structure with alpha helices and beta sheets.



(c) Tertiary structure. The spatial structure consisting of folded helices and planes.



(d) Quaternary structure, representing a protein complex such as in this case *haemoglobin*.

Figure III.1.12: *Protein structure.* The four levels of protein structure.

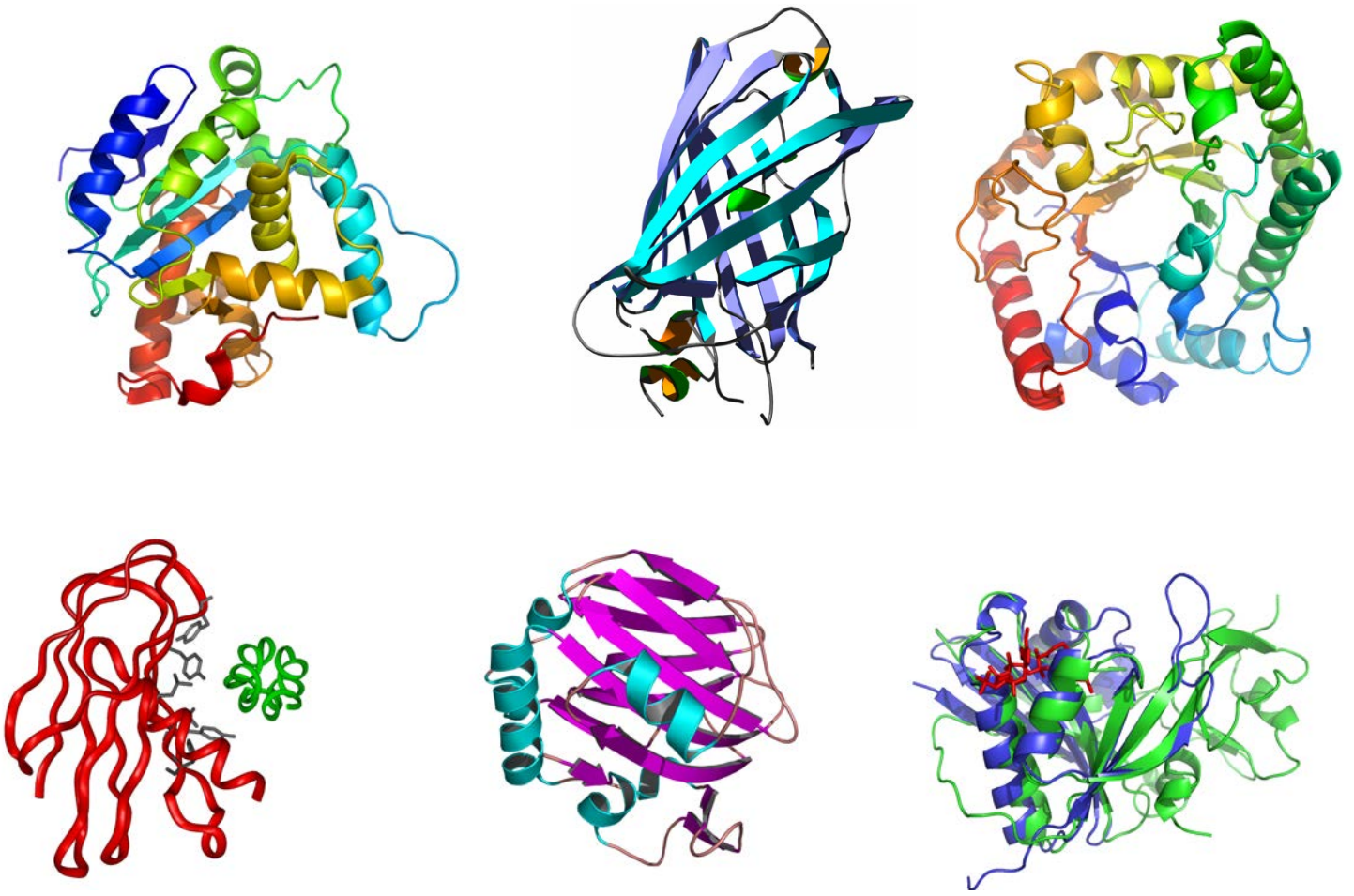


Figure III.1.13: *Proteins: the work horses of life.* Their tertiary three-dimensional structural complexity, diversity and beauty is where the quantum ladder reaches into the heart of life. One could easily imagine trendy fashion designers and hair stylists getting inspiration from these magnificent – all natural – dreadlock designs. For others it is just a splendid paean to reductionism.

stages of cosmic evolution. We have described the wonderful diversity that the flexibility of the carbon atom allows for and that is not only evident in the field of nano-science, but also in biochemistry and molecular biology. We have given examples of how nature has exploited the almost unlimited possibilities to create tremendous diversity from a very limited set of fundamental building blocks.

Further reading.
On molecular physics:

- *Molecular Quantum Mechanics*
  Peter W. Atkins and Ronald S. Friedman
  Oxford University Press (2010)
- *Molecular Physics: Theoretical Principles and Experimental Methods*
  Wolfgang Demtröder
  Wiley (2005)
- *The Molecules of Life*
  John Kuriyan
  Garland Publishers (2012)

Complementary reading:

- *The First Three Minutes: A Modern View of the Origin of the Universe*
  Steven Weinberg
  Basic Books (1977)
- *What is Life?*
  Erwin Schrödinger
  Cambridge University Press (1992)
- *The Double Helix*
  James D. Watson
  Signet Books (1969)

Chapter III.2

The splendid diversity of condensed matter

Water waves are called an emergent phenomenon, because they are a property of the medium water but not of the individual water molecules. Emergent properties, which are ubiquitous in any form of collective, result from the combination of constituent properties and the nature of their interactions.

Condensed states of matter

Condensed matter physics is a research field with a wide scope, because there is a rich diversity of condensed states of matter that we have learned to distinguish and understand. Condensed matter systems are composed of large numbers of constituent particles or agents of various types, each with its own characteristics. When these particles are interacting all kind of unexpected things may happen, and their collective will exhibit a variety of emergent *properties. This raises a question that can be posed in two directions. On the one hand we may start from the observed macroscopic behavior and ask what the microscopic ingredients and mechanisms are that give rise to that collective behavior. On the other hand the microscopic constituents may be given and we are asked to 'design' a 'medium' that exhibits certain macroscopic properties. Condensed matter physics is the systematic study of widely different manifestations of order and disorder. It wants to understand what characterizes the different phases and what the underlying mechanisms are. We start this chapter with an introductory overview of some general concepts and will then focus on specific systems in the following sections. The next chapter is devoted to the properties of the electrons in solids.*

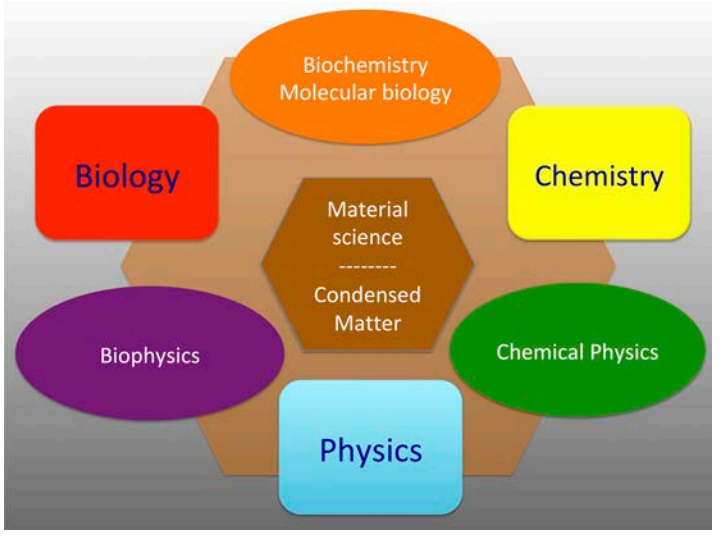


Figure III.2.1: *A science of complexity.* Condensed states of matter are studied in the three basic disciplines, and in the inter- and transdisciplinary fields that emerged in between those disciplines.

A multidisciplinary field. The study of condensed states of matter is by no means an activity only physicists are concerned with. Quite the contrary, it is an inter- or better transdisciplinary field, where the basic disciplines of biology, chemistry and physics, as well as other, interpolating fields, meet and inspire each other in many ways. This research environment is sketched in Figure III.2.1. Generally speaking the understanding of collective – often emergent – behavior, of large numbers of similar constituents or agents, is a principal objective of what is called *complexity science*. But the interactions typically go both ways; from individual to collective and back, from local to global and back. Characteristic for such systems is that they feature a variety of feedback mechanisms whose effects are notoriously hard to understand and model. The models and methodologies developed in statistical physics and condensed matter theory, offer possibilities for adaptation in a much broader context of complexity science – where they have demonstrated to be applicable in disciplines like economics, and other social sciences. Especially with the advent of large-scale computation, which allows large-scale data processing and model simulation (including the nonlinearities representing feedback mechanisms), these parallels can be explored quantitatively.

Just H_2O. Let me start with the familiar example of water. In Figure III.2.2 I have schematically displayed the different phases that can occur as a function of the temperature T. If we start in the middle, say at room temperature, and a normal pressure of one atmosphere, then it will be a liquid. If we heat it, it starts boiling at 100^o C, and will make a transition to the vapor or gas state. And if we cool it, it will freeze and become ice. These phases differ by the way the molecules are aggregated.

Collective behavior. In discussing collective behavior we distinguish a number of conceptual ingredients which we will briefly highlight in this section. On the one hand we have to know what the basic ingredients, often called *constituents* or *agents*, of which the system is composed, are.

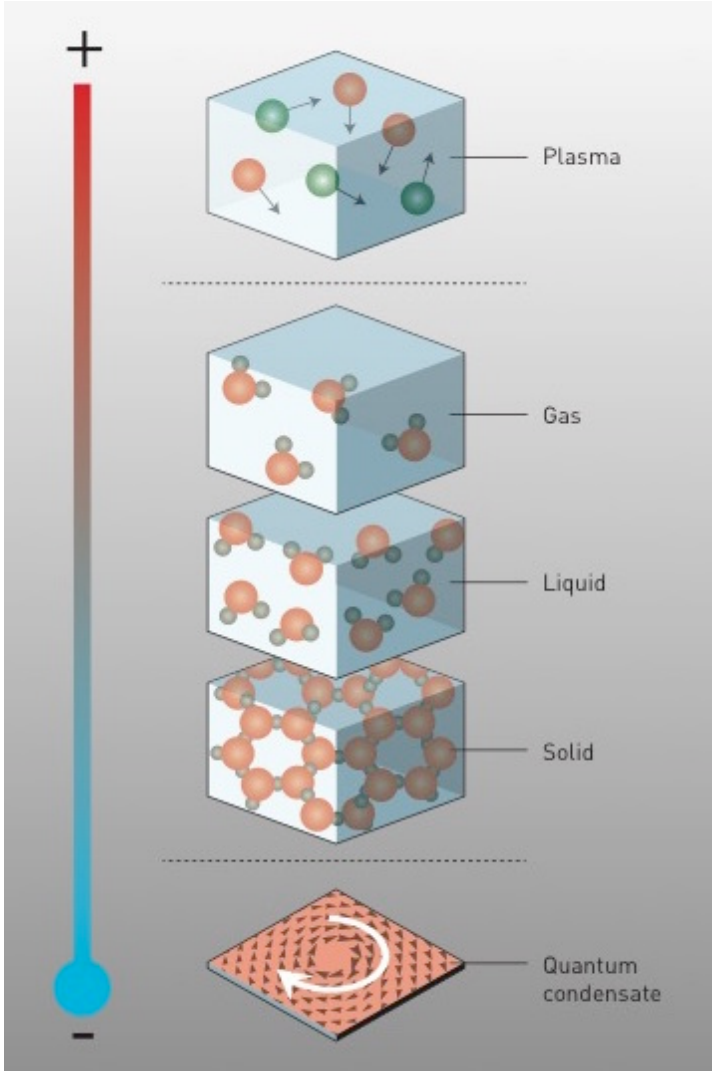


Figure III.2.2: Collective behavior becomes less predictable and harder to understand if we keep lowering the temperature. (Source: Nobel.org)

It is important to know what their individual properties or *internal degrees of freedom* are, but also what their *interactions* look like. On the other hand we have to determine what the possible *external control parameters* are, in the context of physics these are typically things like temperature, pressure and external fields.
The system may have different ways to aggregate, de-

pending on the 'environment' and consequently enter different *phases*. We fix the environmental constraints by choosing the values of the external parameters in certain ranges. These external parameters already refer to macroscopic, that is, collective state variables. The temperature of a gas or liquid for example is linked to the average kinetic energy of the molecules and can be regulated by putting the system in contact with a heat bath.

We are led to the notion of a *phase diagram*, where we draw the space of external parameters (or lower dimensional cross-sections thereof) and divide it into the domains corresponding to the different allowed phases.

Moving through parameter space one encounters boundaries that separate different phases, meaning that the system will go through a *phase transition*. The phases will exhibit different degrees of *order and disorder* on different levels. The question how to distinguish the various phases leads us to the notions of *order parameters* and *correlation functions*.

Finally, once a phase has been recognized, we have to identify the most relevant *effective degrees of freedom* of the system in that phase, these are generally emergent degrees of freedom which do not exist on the constituent level. On the one hand these are the low energy modes corresponding to so-called *quasi-particles*. You may for example think of density waves in a solid which are also called *phonons* or 'particles of sound'. On the other hand in macroscopic media, one often encounters so-called *defects*, these are literally structural defects or imperfections in the medium. Defects can be localized (point like) or extended (like a line or a wall). Defects are robust for topological reasons, and they play a crucial role for understanding the properties of such materials. For example, in a crystal one may have lattice defects, called *dislocations* or *disclinations*, as we will show later on.

Let us now zoom in on the concepts we just introduced.

Constituents and their degrees of freedom. When talking about condensed states of matter, we assume such states to be composed of many constituents. The constituents can themselves be composite as well, like ions, atoms or molecules. The constituents have certain properties like mass, charge, magnetic moments (spins), in fact any of the attributes we have been discussing in previous chapters. The constituents will – depending on their properties – have interactions, and these interactions may be strong or weak, and may be long, short or intermediate ranged. For example, if particles have spin one-half they are fermions and cannot occupy the same state, which has a huge impact on their collective behavior. Relevant is also to what extent the intrinsic degrees of freedom can be manipulated by external controls, like an applied magnetic field for example, which couples to all individual spins in the system. Needless to say that it is precisely the rich variety of constituents and their interactions (including feedbacks) that allow for the splendid diversity of possible states and phases of condensed matter.

In Figure III.2.3 I have indicated the substructures of the most common systems and their typical degrees of freedom which may or may not play a decisive role, depending on the question one is addressing. If we go down in scale the substance may consist of one type or various types of molecules, and much will depend on the *shapes* of the molecules, referring to the charge distributions (the molecular wave functions). These determine the electric and magnetic dipole and higher moments, and as the molecules are overall charge neutral, these moments are crucial and determine the rigidity of the individual shapes. And clearly these shapes are all-important for understanding how the molecules can fit together in a stable way, which in turn determines the allowed symmetries of a crystal to be formed. If the molecules become large, like polymers for example, one can imagine complex materials being assembled, like biological tissues made from large biomolecules.

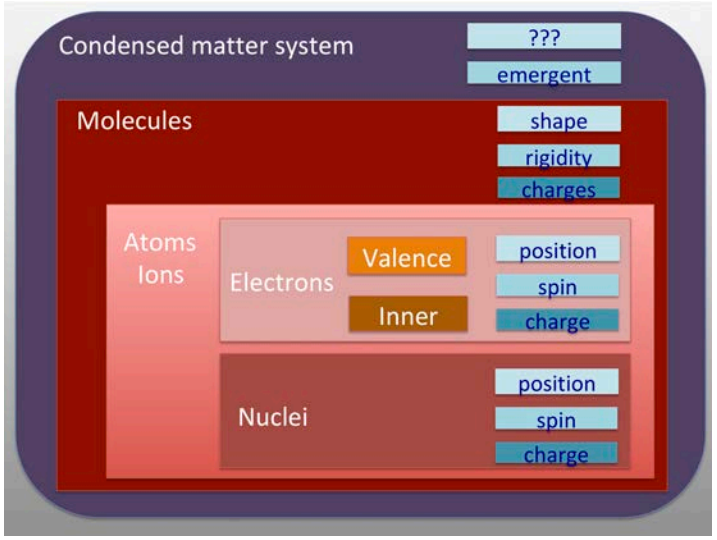


Figure III.2.3: *A hierarchy of degrees of freedom.* Building blocks of a condensed matter system (in white) and their 'degrees of freedom' (in blue).

The relevant constituents may also just be atoms, and they may form crystals, where they optimally balance their kinetic and potential energy, or alternatively their attraction and repulsion. The picture is that the nuclei sit on the sites of a lattice and the electron states may either be localized on the nuclei or be spread out and extended. The electrons in the outer shell – so-called valence electrons – are relatively weakly bound and can hop to neighboring sites and in the case we are dealing with a conductor, they even have non-localized states that spread out over the whole lattice. So, the material is a highly ordered solid, but hidden in there are the electrons which form a freely streaming (not-ordered) fluid supported by the solid substrate of highly localized ions. Similarly, we may have a solid where, say, the atomic spins are ordered, in which case we have a ferro or anti-ferromagnet, or the spins may be disordered – pointing in random directions – and there would be no overall magnetization. And indeed, these charge and spin degrees of freedom can be manipulated by imposing external electric or magnetic fields.

Control parameters and phase diagrams.

An important remark is that the 'relevant degrees of freedom' of the system as a whole are not known *a priori*, exactly because they will mostly be emergent such as sound, spin waves, currents, defects etc. These emergent degrees of freedom will strongly depend on the choices we make for the external parameters. These are for example the thermodynamic parameters such as temperature, pressure or chemical potential. Other parameters correspond to external electric and magnetic fields, or the chemical composition (or doping) of the material. Moreover, there is a dependence on the dynamic of preparation. If we cool a liquid rapidly (called *quenching*), then it may not have had enough time to achieve the optimal type of long-range order. It would stay somewhat amorphous, in contrast with the perfect crystal which forms if we cool the liquid down slowly (called *annealing*).

There are still other options for manipulating the system. You may change the relative concentrations of components. You may replace certain components by similar, or not so similar ones. You can add components (like *solvents* or *interstitials*), or 'dope' the system by adding or removing charge carriers. These tools have been used in the most inventive ways to engineer materials with specific, sometimes most unusual, but highly desirable properties. This advanced form of 'legoism' makes certain corners of material science look like a kind of black magic: a form of witchcraft with the distinctive feature that it works!

The phase diagram. The parameter space may be divided into domains corresponding to the different phases, and this information is usually represented in a *phase diagram*. Often we are interested only in particular phenomena and we can restrict ourselves to smaller- and lower-dimensional cross sections of the parameter space. One axis that is usually present is the temperature (or energy) axis, and another is for example the pressure (or density) axis. If we add the pressure P, we can extend the Fig-

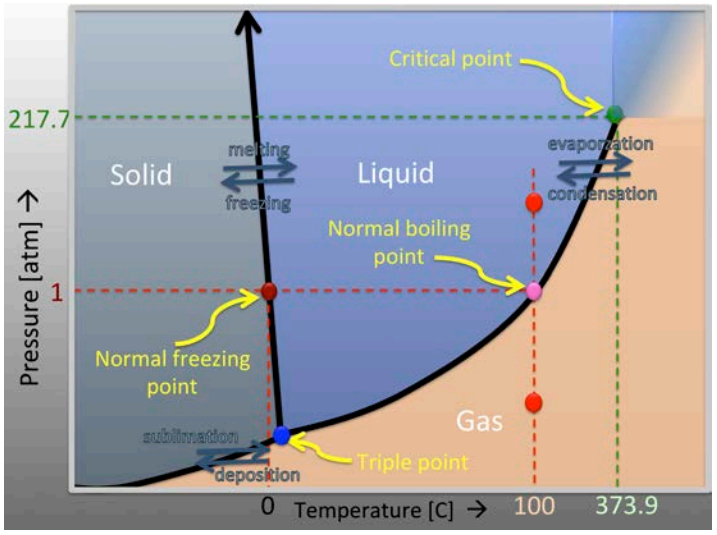


Figure III.2.4: *Phase diagram.* The standard phase diagram of ice/water/vapor with the triple point, and the standard definition of boiling and freezing point. Above the critical point there is a smooth crossover from the liquid to the vapor state.

ure III.2.2 to the two-dimensional $\mathrm{P}-\mathrm{T}$ phase diagram of Figure III.2.4. This adds novel features: the normal boiling and freezing points become lines, and as we see, these lines may join (or split) at a so-called *triple point*. Furthermore a line may terminate at a so-called *critical point*, where a clear distinction between the phases ceases to exist.

Equation of state. The state variables are usually not independent, since they have to satisfy a constraint, which is called the *equation of state*. For a fixed amount of stuff, say one *mol*, which means a total number of $\mathrm{N_A}$ molecules, one finds that in the diluted gas phase for example the 'ideal gas law' holds. This law states that $\mathrm{PV}=\mathrm{RT}$, which is a functional constraint on the macroscopic state variables P, V and T involving the universal or molar gas constant $\mathrm{R}=\mathrm{N_A k}$ which is just a fixed number (the product of Avogadro's and Boltzmann's constants). If we for example consider a fixed amount of gas in a container of fixed volume V, the equation tells us that lowering the temperature would lower the pressure proportionally (at least in a lower right-hand side of the diagram where the 'law' holds).

Phase transitions. Crossing a phase boundary in a phase diagram means that the system goes through a *phase transition*. Let us for a moment look at the dark blue line separating the liquid and gas or vapor phases. Crossing that line from blue to light brown means boiling the liquid. What you immediately see is that this may happen on any point on that line segment. If we boil an egg on a Sunday morning, what we do is that we have a fixed normal pressure of 1 atmosphere, and by heating the water we move to the right on the dashed red line until we hit the transition point at 100 degrees Celsius. But a less practical way to boil an egg would be start at high pressure with water at $100^{\mathrm{o}}\ \mathrm{C}$, the water is not boiling then but when we lower the pressure, sure enough when it hits 1 atmosphere the water would start boiling. This boiling process would correspond to crossing the phase boundary top down along the vertical dashed red line starting at the high red point moving to the pink straight below. High in the mountains the pressure of the atmosphere is lower and thus water boils at a lower temperature (about 4 degrees per kilometer elevation), which can make preparing your soft-boiled Sunday morning egg quite a hassle. Often phase transitions signal the occurrence of a tipping point in some (free) energy landscape of the system due to changes in the control parameters. And in that sense the phase diagram is a natural characterization for any multi-particle or multi-agent system.

Critical points. In a critical point, a phase separation line terminates. This means that the clear distinction between the two phases, and the marked transition between them, somehow disappears. We enter a critical region in which there is a smooth crossover between – in this case – the liquid and the vapor. In fact, the usual clear surface separating them disappears and becomes a foggy layer.

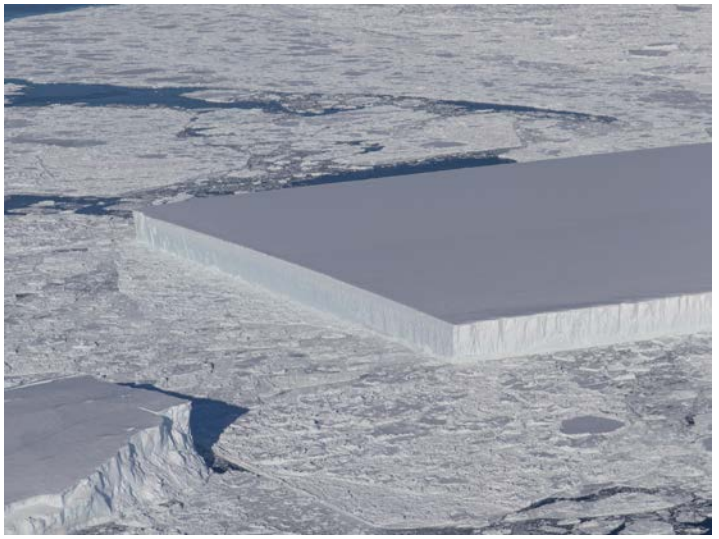


Figure III.2.5: *A tabular iceberg.* In October 2018 a NASA inspection team discovered this huge, perfectly rectangular, so-called tabular iceberg in the arctic. Such bergs are formed naturally and have a strikingly rectangular geometry, reflecting the underlying crystal structure. They are not single giant monocrystals, though they look like it. (Source: NASA ICE)

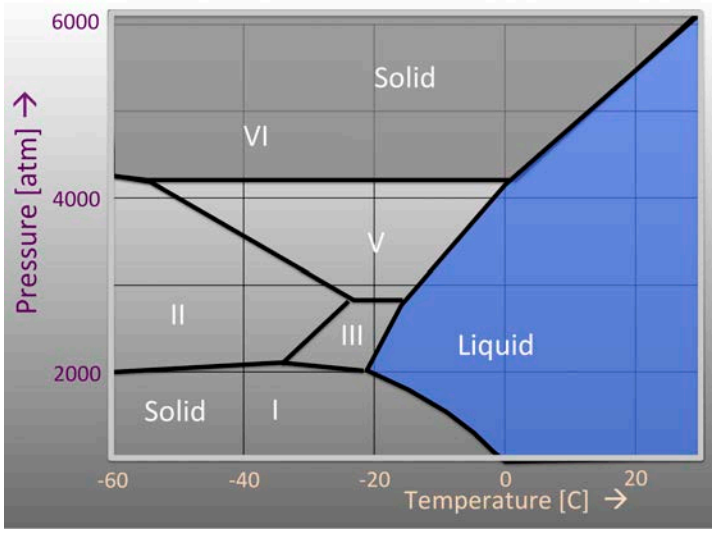


Figure III.2.6: *Ice varieties.* In the high-pressure regime, high up along the vertical axis of the previous diagram, there are many distinct solid phases of water, where the water molecules happen to organize according to different symmetries.

Ice? What ice? In Figure III.2.6 we show a tiny corner of the phase diagram of water at very high pressures, and therefore not present in Figure III.2.4. It would have appeared high up on the left, in the direction where the arrow is pointing. The diagram shows that if you make the pressure large enough, the water will become solid even at higher temperatures. You furthermore see that there are actually many distinct solid phases up there. They are forms of ice that differ by their crystal symmetries. Some are *hexagonal* (I) others *tetragonal* (III, VI), *monoclinic* (V), *rhombohedral* (II) or *cubic* (not in the graph). A true *Baskin & Robbins* of structures, but – I am sure – all equally tasteless. Furthermore, many of these fancy phases are metastable, so they tend to decay in more stable versions. Note also the impressive number of triple points in the phase diagram. Such is the hidden diversity of something as common as water. It shows its complex behavior only under extreme conditions.

Water versus Argon. It is interesting to compare the features of the phase diagram of water, with for example that of the noble element argon. The element $^{36}\mathrm{Ar}$ has 18 protons, and its 18 electrons completely fill all energy levels up to the $\mathrm{n} = 3,\ \mathrm{l} = 1$ shell of atomic states. These completely filled shells make the element stable and resistant to bonding to any companion. The noble elements are 'Einzelgängers', or 'lonely cowboys' so to say, they apparently have everything they need, and are like extreme individualists who love to ignore their neighbors. Under normal conditions it is an inert gas, and it has a phase diagram similar to that of water as depicted in Figure III.2.4, though the corresponding points are positioned at different locations. As is clear from Table III.2.1, for argon things happen at much lower temperatures, which indeed is a consequence of their 'nobility.'

If you would continue the phase diagram for argon to high pressures, one would surely see the melting line bend over to higher temperatures, meaning that liquid argon would

Table III.2.1: Comparison of water and argon.

Phase diagram	Water		Argon	
	T[K]	P[atm]	T[K]	P[atm]
Melting point	273.15	1	83.81	1
Boiling point	373.15	1	87.30	1
Triple point	273.16	0.006	83.81	0.68
Critical point	647.10	217.7	150.69	48.0

just like water solidify at very high pressures. This, however, only happens at pressures of tens of thousands of atmospheres! Furthermore, because this simple atom has so few degrees of freedom, it exhibits only one solid phase. This means that the phase diagram III.2.6 for argon would be rather boring, because it would just show one melting line going across from left to right.

At this point two observations can be made. On the one hand there are universals in phase diagrams, like that condensed matter will become solid under high pressure or at low temperatures (including the familiar triple and critical points). On the other hand, phase diagrams may exhibit a huge structural diversity that depends on the specifics of the constituents, whether they are simple spherical atoms, or composites with many internal degrees of freedom like water molecules.

Crystals. The reason that solids – usually crystals – form is that by bringing many atoms close together the orbits of the electrons start overlapping and the electrons start moving around changing nuclear partner so to say, which leads to an effective attraction. However, if they get too close the effect of the repulsion of positively charged nuclei starts to dominate. Balancing attraction and repulsion the atoms tend to organize themselves into an optimal pattern that minimizes their overall interaction energy. This is basically how crystals form. In a crystal the positioning of the atoms is strictly periodic which implies strong spatial correlations over large distances, corresponding to some discrete translational (and rotational) symmetries. Complexity and beauty apparently arise where attraction and repulsion strike a subtle balance.

Hard versus soft condensed matter. The field of condensed matter physics is divided up into two parts: soft and hard condensed matter physics comprising the topics we have indicated in Figure III.2.7.

Soft matter. With soft matter we think of liquids, coloids, gels, molecular materials like polymers and biomaterials. It is a diverse field that often involves physics at an intermediate – so-called mesoscopic – scale, like nano structures for example. This field mostly employs methods from classical physics, such as statistical mechanics and classical field theory, but also lots of chemistry. It is the branch of condensed matter physics most remote from hard core quantum theory, but it has become an innovative field with a wide range of applications. One of its most influential protagonists was Pierre-Gilles de Gennes of the École Normale Supérieure at Paris, who received the 1991 Nobel prize for his extensive oeuvre. This field has led to beautiful insights into the role of symmetry and its breaking. We will therefore in the following not just discuss crystals, but also *liquid crystals* and *quasicrystals*.

Hard matter. Hard matter is the present incarnation of what used to be called *solid state physics*. It studies properties of materials where quantum theory is absolutely indispensable. Quantum properties are vital for understanding the role electrons and lattice vibrations play. In the quantum realm these can rearrange themselves in collective quasi-particle degrees of freedom, with totally unexpected emergent properties, like *superfluidity*, low and high temperature *superconductivity*, and *topological order*. These latter phases, for example fractional quantum Hall systems, include new degrees of freedom called *anyons* with exotic spin and statistics properties. Towards the end of Chapter III.3 we will take a closer look at them.

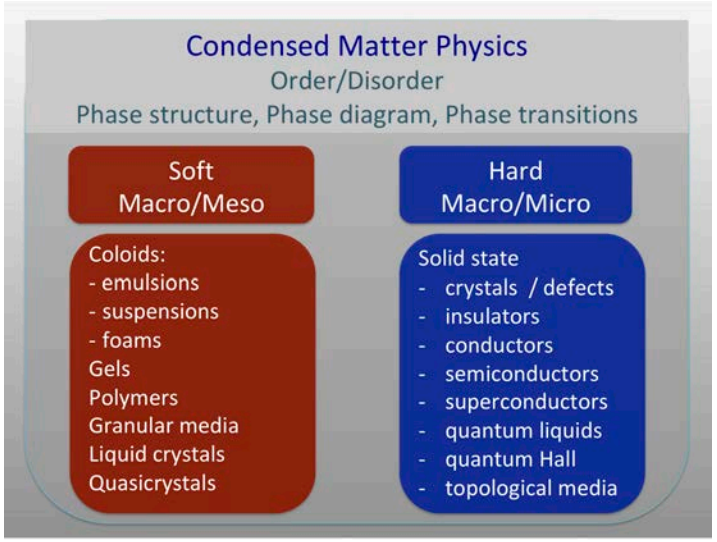


Figure III.2.7: *Hard versus soft.* Condensed matter can be roughly divided up into 'soft' and 'hard' matter. Both are of great technological importance.

Plasma. We have seen that for large pressures most systems become solid. There is of course also the other extreme regime, corresponding to high temperatures, which is of interest as was already indicated in Figure III.2.2. For very high temperatures, there is yet another phase transition: the water molecules will ionize, which means that they will break up in two oppositely charged components, the OH^- and H^+ ions:

$$H_2O \rightarrow OH^- + H^+$$

This is again a quite different state of water. It is still overall electrically neutral, but it will couple strongly to electric and magnetic fields, because the individual components (and constituents) do. If you apply a voltage over the plasma, currents will flow, and clearly, the positive and negatively charged components will run in opposite directions.

In Chapter I.3 where we talked about fusion, we mentioned the crucial role played by the tritium plasma as a 'fuel'. And in the previous chapter we alluded to the state of the very early universe as a *primordial soup*, this refers to a universal plasma made up of bare 'charges' for all interaction types. Of special interest is the colored component of the soup called the *quark-gluon plasma*, which is nowadays studied experimentally by smashing lead ions into each other in the Large Hadron Collider at CERN, by the so-called ALICE collaboration. In that experiment one tries to recreate for a tiny period of time, a tiny bit of early universe. It is fascinating to realize that not only with space observatories but also with big accelerators one is trying to get ever closer to the Big Bang and thus contributing to cosmology.

Order versus disorder

We have indicated the importance of identifying different phases. These are roughly characterized as ordered and disordered phases, but also phases that sit in between. Solids are highly ordered, gases are disordered, and simple liquids tend to be more like dense gases, but if the constituents are more complicated, they can be both. Both ordered and disordered! How can that be? Well, it depends on which degrees of freedom you are talking about. In a liquid crystal for example, the positions of the molecules are not frozen into a crystal (disorder), but the orientations of the molecules are all aligned (order). Glass appears to be solid but is in fact an extremely viscous liquid. And what about gels, polymers, and biomaterials, are they ordered and in what ways? In a conductor the nuclei have fixed positions in the crystal lattice, yet at the same time the conducting electrons form a liquid that flows freely through the material. The diverse topics we have mentioned so far used to belong to different fields of study but are more and more integrated because similar techniques are used to study them.

One of the fascinating results from classical physics, in particular statistical thermodynamics, is that certain disordered equilibrium states like a gas of atoms or a liquid

can still be rather easily described if one applies statistical methods to them. After all, the behavior of a *mol* of a dilute gas consisting of some 10^{23} atoms in equilibrium, can to a first approximation be described in terms of only a few macroscopic variables like pressure P, temperature T and a volume V and an entropy S, that have to satisfy the *ideal gas law*, $\mathrm{PV} = \mathrm{RT}$. Such a drastic reduction of variables can be performed if one is only interested in the most relevant degrees of freedom that effectively describe the equilibrium states of the collective in a given phase.

The gas molecules bounce around randomly, yet though the individual behavior of the atoms is highly erratic, the collective is surprisingly well behaved and predictable. As every insurance company can tell you, if the number of clients is sufficiently large, statistics becomes an extremely reliable tool for predicting the probability of certain events. In the classical theory of somewhat less diluted gasses, where one takes the size of the atoms and the presence of walls of the container into account, one arrives at the *Van der Waals equation of state*. This equation is an important generalization of the ideal gas law from a conceptual point of view, because it predicts a phase transition to a liquid state. We will return to this equation shortly.

It turns out that the most complicated behavior is observed near a phase transition. There the distribution of thermal fluctuations broadens; fluctuations apparently occur on all scales which means that they are not distributed like a Gaussian distribution with a well-defined mean and variance around the mean. No, the distributions behave like *power laws*, where compared with the Gaussian, the venom is in the tail of the distribution. Whereas the exponential distribution tends rapidly to zero, the power laws have so-called *fat tails*. These tails describe so-called 'high impact, low probability' events, but the point is rather that although that these events are far away from the average, their probability is actually *not* so small after all, in fact gigantic compared to an exponential distribution With power laws extreme events in the tail of the distribution cannot be discarded at all. Indeed, under such circumstances, insurance brokers are not that eager anymore to sell you an insurance policy, and if they do, they will certainly make you pay a good deal more to cover their substantial risks.

Phases, order parameters and correlations. So what then determines in what sense a system is ordered or disordered?

Order parameters. There is a special set of observables important for the identification of different phases: these are denoted as local *order parameters*, which are called local because they depend on the position x. To probe the difference between a vapor and its liquid state, the order parameter would be the local density $\rho(\mathrm{x})$. In the transition it would make a sudden jump from a tiny to a large constant value $\rho(\mathrm{x}) = \rho_0$. For magnetic systems the order parameter is the magnetization M, which is the spatial average of the local magnetization $\mathrm{M}(\mathrm{x})$, which in turn corresponds to a local average of a sizeable number of spins centered around the point x. In metals spontaneous magnetization occurs at the so-called Curie temperature, which means that the magnetization M acquires a non-zero value below this temperature. So, to conclude, order parameters are specific observables that probe for a structural change in the state of the system when it goes through a phase transition.

First- and second-order phase transitions. We distinguish two types of phase transitions called first- and second-order transitions. For the second order transition the order parameter changes continuously (but not smoothly) from zero to a non-zero value. A typical example is spontaneous magnetization which we just mentioned and will discuss in more detail shortly.

Correlations. The order parameters correspond to the average property of a local quantity. But a measure of order can also be more subtle and correspond to probing multi-

local correlations in space or time in the system. For example, if you have a crystal, then many of its properties are periodic, and strongly correlated spatially. Measuring such correlations can then help you identify the spatial structure and symmetries of the crystal. A famous technique, X-ray diffraction, does exactly that: it yields a diffraction pattern in which such spatial correlations are encoded, and from which the three-dimensional crystal structure can be reconstructed.

First-order transitions. In a first-order transition the order parameter jumps discontinuously. A nice example is the liquid-vapor transition (evaporation or boiling), where there is a region in parameter space where both phases can coexist, but where one of them becomes unstable. The transition then often takes place through bubble nucleation, as we know too well from the ordinary boiling phenomenon. Inside the bubbles we have the new phase and outside the bubble is still the old liquid phase. Because of thermal fluctuations, bubbles spontaneously form in the liquid, and if they have a sufficient size they will start growing. The threshold occurs when the energy it costs to make the wall (proportional to the surface area of the bubble) becomes equal to the energy gain which is given by the energy difference between the two phases, and this gain is proportional to the volume of the bubble. Clearly, if the bubble is large enough the volume term wins, and the bubble will start expanding. If you transfer more heat to the liquid, more and larger bubbles will form, and those may furthermore coalesce. This process continues until the transition is completed and there is no fluid left.

In the Figures III.2.8 and III.2.9 we have depicted the liquid-vapor transition from two complementary points of view. The first figure shows the transition in a pressure-volume (P, V) diagram. The colored curves are different isotherms (curves of constant temperature). The yellow one corresponds to a high temperature and reproduces the ideal gas law, $\mathrm{P} = \mathrm{RT}/V$. The orange isotherm where $\mathrm{T} = \mathrm{T_c}$ is special because all lower isotherms have a minimum

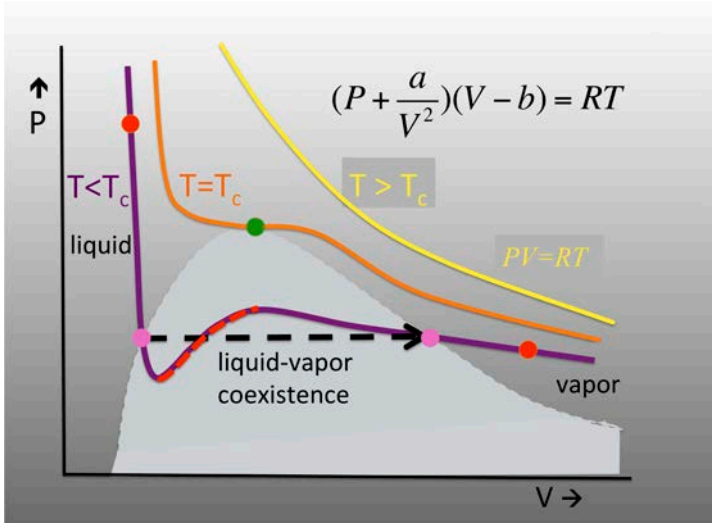


Figure III.2.8: *Van der Waals equation of state.* We have sketched three isotherms meaning P as function of V with T fixed. The yellow one for $\mathrm{T} > \mathrm{T_c}$, where we recover the ideal gas law. The orange one is for $\mathrm{T} = \mathrm{T_c}$, and the purple one corresponds to the boiling process as described in the text.

and a maximum. The purple curve is the 100^{o} Celsius isotherm and describes the process corresponding to the vertical transition marked in Figure III.2.4. The points on an isotherm supposedly correspond to equilibrium states, but that cannot always be the case. The segment highlighted with the dashed red line cannot represent physically acceptable states because increasing the volume would also increase the pressure, but for physical states it is the other way around, the 'compressibility' in those points has the wrong sign. So only the descending parts of the isotherm represent allowed equilibrium states. What makes these curves interesting is precisely that for $\mathrm{T} < \mathrm{T_c}$, we see that for a certain pressure range there are two possible states: the left one corresponding to the liquid and the one on the right to the vapor. The picture does immediately suggest the explanation. We can slowly descend the 100^{o} isotherm by increasing the volume and thereby lowering the pressure, keeping the system in equilibrium until we hit the dotted line at the pink point (where $\mathrm{P} = 1\ \mathrm{atm}$). This is where

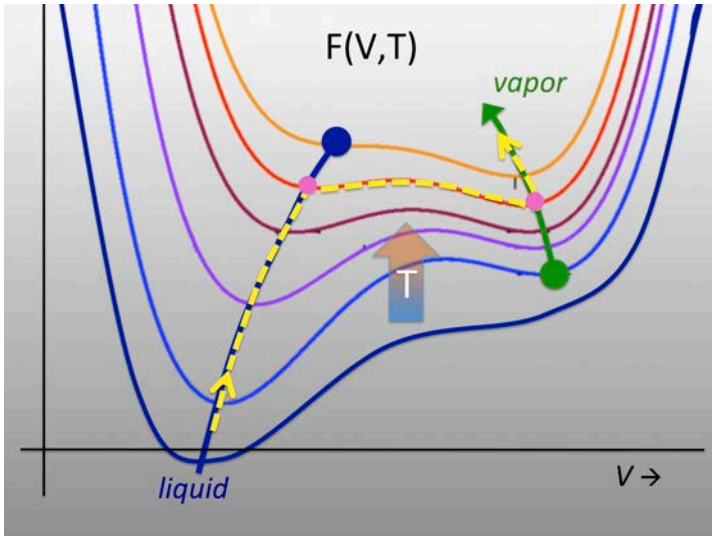


Figure III.2.9: *Free energy landscapes for different temperatures.* The minima correspond to the equilibrium liquid and/or vapor states. The yellow trajectory corresponds to horizontal 'boiling' trajectory in the phase diagram of Figure III.2.4. The liquid state is stable, until we hit the red curve where the vapor minimum has lower energy and the system makes the boiling transition to that stable vapor state.

the boiling transition starts, and as we all know this is a pretty violent non-equilibrium type of process that works through bubble nucleation and continues until all the liquid has vaporized and the system can restore equilibrium in the vapor state on the isotherm (corresponding to the pink point on the right). From there the system may move down again if the volume is further enlarged. In the intermediate region during the crossing we have two coexisting phases in the system, part is liquid and part is vapor. The whole transition trajectory marked by the dashed black arrow between the two pink dots, thus corresponds to the single pink dot in the phase diagram III.2.4. This teaches us that the phase diagram certainly tells us that there is a transition but does not inform us in any way about how that transition actually takes place, and whether it is a first- or second-order transition.

Minimizing the free energy. Now in the second figure, Figure III.2.9, we look at the first-order transition from the point of view of the free energy $\mathrm{F} = \mathrm{F}(\mathrm{V},\mathrm{T})$ of the system, and this time it is convenient to take the horizontal trajectory in the phase diagram, corresponding to the familiar boiling process we witness in the kitchen.[1] In the figure we plotted the free energy as a function of volume for increasing temperatures. The equilibrium states correspond to minima of the free energy and we see that there is a range of temperatures where we have two minima. We have a fixed amount of matter, so the left minimum is the small volume or liquid state, and the right minimum is the vapor state. We start at a low temperature equilibrium state corresponding to the unique minimum. If we start raising the temperature, we see that the energy landscape is changing. Once we arrive at the light blue isobar it develops a second (local) minimum, but it has higher energy and is therefore unstable. If an outlandish fluctuation somewhere in the liquid happens to create a tiny vapor bubble, this bubble would instantly collapse because there is nothing to gain (energy-wise) by being a bubble. However by going to higher temperatures the values of F for the two minima become equal, and on the red curve the vapor minimum has become clearly lower than the liquid one. Then indeed, the liquid state becomes metastable. Even moderate fluctuations will create bubbles that are big enough to start growing, thereby executing the actual vaporization process. You also see that even if we are careful and succeed in overheating the liquid, then you hit the dark blue point where the minimum corresponding to the liquid disappears. At that point the liquid state becomes unstable and the transition necessarily takes place.

Tipping points. It is worth pointing out that the free energy diagram is quite universal for understanding the origin of tipping points in all kinds of multi-agent systems. The free energy would correspond to some relevant 'util-

[1] This is a process at fixed pressure, and is naturally presented by equal pressure lines or so-called *isobars* in a (T,V) diagram.

ity function' the system wants to minimize (environmental constraints, costs, etc). The 'fitness' landscape will in general depend on other (control) variables. For example, we have a society burning fossil fuels which provides us energy for $\$X$ per kWh. Around 1970 the landscape started to change in that another possibility appeared, namely solar power. It is still expensive, and without some local subsidies a local effort easily collapses. However the price comes down rapidly, and the second minimum of the utility function starts competing. Ambitious countries, states and cities may create successful local bubbles that are economically feasible and start to grow. And that is how the energy transition will presumably take place in the present age. The energy transition is typically a first-order transition, and as a matter of fact we see it happening all around us! This shows you the metaphorical power of the boiling process as a model for certain types of transitions and the visualization with the two competing minima as a powerful analogy.

Collective degrees of freedom: quasi particles. Once the system has chosen a different ground state corresponding to a new phase and another minimum of the free energy, we should ask what other aspects of the physics of the system have changed. Most importantly, we should find out what the low energy excitations of the system in the new state are. The low energy excitations are of interest because they are the first that will get excited if we perturb the system, and as such they determine more then anything else the *emergent* properties of the system in the new phase. These modes help also to identify and label the collective states. Whether it is a conductor to heat or electricity, or whether it is a magnetically ordered ferromagnet, for example.

What happens to a crystal if I hit it? This is like probing the system by locally deforming it and observing the response of the system to that deformation. We study how the deformation propagates through the system. How the deformation energy starts spreading. The resulting propagating modes are the low energy excitations, in this case they are longitudinal density waves, which correspond to sound. Sound is an emergent phenomenon because an individual atom does not know what sound is, it cannot make sound by itself. It needs the ordered collective to propagate, and in that sense it is just like the 'wave' that can be excited in a football stadium: to let it propagate through the crowd requires a collective effort. And if a large fraction of the audience are fans of the opposing team, it will definitely not propagate. The point I am making is that by studying the response of the system to perturbations we get to know a lot about its ground state or phase.

In reality the molecular systems we consider are more omplicated and we do not only have to worry about the positions of the nuclei in the crystal lattice. For example, the nuclei may have a tiny magnetic moment, called *spin*, which means that they are like tiny bar magnets. If the system is at a relatively high temperature these little magnets will point in arbitrary directions. They are highly independent, and thus their orientations are uncorrelated even on short distances. So, in this case we have that the nuclei are strictly ordered because they form a crystal, while their spins are not ordered at all. Apparently, we have to be specific if we say that a system is ordered.

The behavior of electrons. Another crucial ingredient of most condensed matter systems that we have not mentioned so far are the electrons. Given the underlying lattice structure of the nuclei, what is the quantum behavior of the electrons in that given background? Do they stay localized, close to 'their' nucleus, or do they start hopping around freely, or do they form a conducting fluid of some sort? It turns out that the behavior of the collective of electrons in condensed states of matter is highly diverse and keeps surprising us up to today. Understanding this behavioral variety is one of the main drivers of condensed matter physics. These problems have been studied for decades and time and again new fundamental properties are discovered often leading to important technological in-

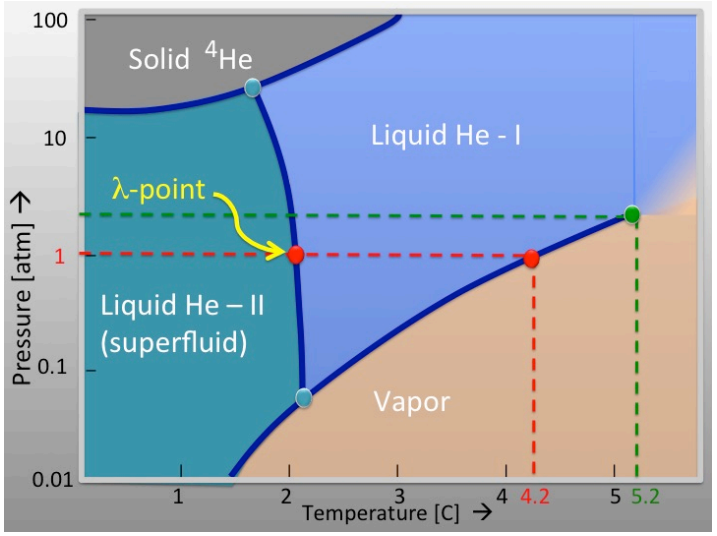


Figure III.2.10: *Phase diagram of* $^4\mathrm{He}$. Comparing this phase diagram with the conventional one of Figure III.2.4, a new superfluid phase has opened up at low temperature, splitting the triple point into two triple points. The critical point is marked in green.

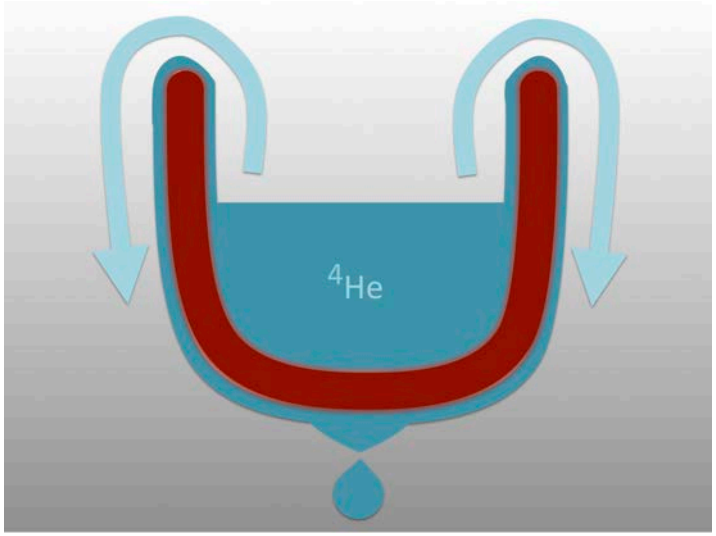


Figure III.2.11: *Superfluidity.* The vessel filled with liquid $^4\mathrm{He}$ that will turn into a frictionless superfluid. When cooled below the $\lambda-$point, it will spontaneously creep over the wall of the vessel until it is empty.

novations. We focus on some of these types of behavior in the remainder of this section. A more thorough analysis is given in the next chapter.

The quantum regime. In Chapter I.3 we discussed scales and units, and pointed out that at low energies quantum theory necessarily comes into play, which leads to another plethora of conceivable physical states that can have highly unusual properties like superfluidity and superconductivity.

In the quantum regime we should expect that the quant- essential spin and statistics properties of particles come into play but also that the Heisenberg uncertainty relations will manifest themselves in the collective behavior. Of special interest is the possibility that bosons can occupy the same state. What typically happens is that once you lower the temperature far enough, a macroscopic number of the bosonic particles will occupy the same lowest energy state. The system forms a so-called *Bose condensate*, a special quantum coherent state, which means that the system will go through a phase transition. Systems where this happens will exhibit 'macroscopic quantum' behavior. Quantum matter phases have been in the centre of attention for quite a long time, and still many novel phases are discovered, which pose formidable puzzles for the theorists to understand, like for example high temperature superconductivity. There are still many open questions with regard to understanding collective quantum phenomena from a microscopic, first principles point of view.

Superfluidity. Let us consider the famous example of Helium-4, a boson, where you can see how the quantum behavior, the formation of a Bose condensate, adds a new phase to the phase diagram. The phase diagram of Figure III.2.10 shows the actually not so recent discovery of *superfluidity* by the Russian physicist Pjotr Kapitza in 1937 (and independently by J.F. Allen and D. Misener).[2]

[2]The discovery was made at a time of international tensions, and therefore credentials have been somewhat controversial. An interest-

He received the 1978 Physics Nobel prize for his landmark contributions to low temperature physics, of which this discovery clearly was an outstanding one. We see that compared to the standard phase diagram of Figure III.2.4, in the low temperature region the superfluid phase has been added. Kapitza discovered that phase by just lowering the temperature of some $^4\mathrm{He}$ vapor under the standard pressure of one atmosphere, so he came from the right at the height of the horizontal red dotted line in the diagram. He first crossed the 'standard' transition from vapor to fluid, but then at a temperature of 2.17 K at the so-called λ point he witnessed the transition to the superfluid phase. The ordinary triple point had 'opened up' and with the appearance of the new phase it split up into two triple points. A superfluid displays the curious property of frictionless flow, and therefore behaves rather 'creepy' in the literal sense. If you watch an open container filled with superfluid, you will see the fluid all by itself creep over the rim and run down the outside of the vessel. In Figure III.2.11 we have sketched an experiment along these lines: the self emptying mug! Thank heaven there is friction! Thank heaven that our superdrinks are not superfluids!

Magnetic order

Magnetization. Magnetic properties of atoms are the combined result of three components: (i) the electrons have spin with an associated magnetic moment of one Bohr magneton μ_b; (ii) the atomic orbits of electrons correspond to states with a magnetic quantum number m, which means that the magnetic moment of the orbit equals $\mathrm{m}\mu_\mathrm{b}$; and (iii) finally there is the nuclear magnetic moment which turns out to be a factor thousand smaller. We will not enter in any detailed discussion of how these interact but will just assume atoms, ions, or electrons to have some overall spin or magnetic moment. For the spins we can now also introduce an order parameter, it is called the *magnetization* $\mathbf{M}(\mathrm{x})$, the average magnetic orientation of certain number of spins around the point x. If the temperature is high we know that because of the random orientation of the spins, the average magnetization $\langle\mathbf{M}(\mathrm{x})\rangle$ in the ground state will be zero. But if we cool the medium down, then the disturbances in the lattice become smaller and the magnets will feel each other and can lower the energy of the state by aligning, in which case a phase transition will take place.

Phase transition at the Curie point. At a certain temperature called the Curie point there will be a *phase transition* to a state where all spins will spontaneously align. Order is spontaneously created and the order parameter will acquire a non-vanishing constant, that is to say a position independent value: $\langle\mathbf{M}(\mathrm{x})\rangle = \mathbf{m}_0 \neq 0$. This emergent form of order is called *spontaneous magnetization*, and the system is in a ferromagnetic phase and as a whole behaves like a single big magnet. So, we may conclude that ordinary permanent magnets are made of materials of which the Curie temperature lies far above room temperature. And as expected the order parameter thus signals whether the system is ordered or disordered.

Low energy modes: spinwaves. The low energy modes associated with the magnetic spins in a ferromagnet are the so-called *spin waves*. You may compare them to the waves that a light breeze can excite in a field of grain as we described in the section on symmetry breaking in Chapter II.6. These are again collective excitations of the ordered spin system with a wavelength that is long compared to the distances between the spins and because they have a long-wavelength they are low energy excitations indeed. If we quantize these waves, we get particle like excitations or quasi-particles called *magnons*.

ing historical account can be found in *S. Balibar, The discovery of superfluidity, Journal of Low Temperature Physics, Vol. 146, Nos. 5/6, 2007.*

The Ising model

Let us take some time to discuss an truly iconic model that instantly comes to the mind of any physicist when you mention the word phase transition. It is called the *Ising model*, cherished for its simplicity and its depth, which was introduced by Wilhelm Lenz in 1920. He suggested it as a problem to his student Enst Ising, who then solved the one-dimensional version of it and found that it had *no* phase transition. Moreover, he erroneously concluded that there would be no phase transition in any dimension. How ironic that Ising's 'fame' in physics is based on drawing a wrong conclusion from an elementary calculation. It is precisely that two-dimensional version we are going to discuss, which has for a long time been the canonical model for a (second-order) phase transition. It was solved exactly by Lars Onsager in 1944, who reportedly at a conference just wrote down the exact answers on a black board without further explanation, leaving the learned audience flabbergasted, and with a nice problem to work on! The problem of figuring out how he did it. It is one of those models to which a tremendous amount of work has been devoted. It has popped up in all subfields of physics and beyond.

As mentioned before, we distinguish the ordered *ferromagnetic phase* where all spins are aligned, and the *non-magnetic phase*, where the spins point in random directions. Here the order does not concern the spatial positioning but the orientation of the spins. As we pointed out, in the ordered phase the magnetization is some non-zero constant while in the disordered phase it is equal to zero. To be precise there is a different ordered phase which is called anti-ferromagnetic, where the spins at neighboring sites are anti-aligned.

The Ising Hamiltomian. The classical Ising model has an *infinite array of spins that can only point up or down*. A two-dimensional Ising model configuration is depicted in Figure III.2.12. The spins $\sigma_i = \pm 1$ only interact with their

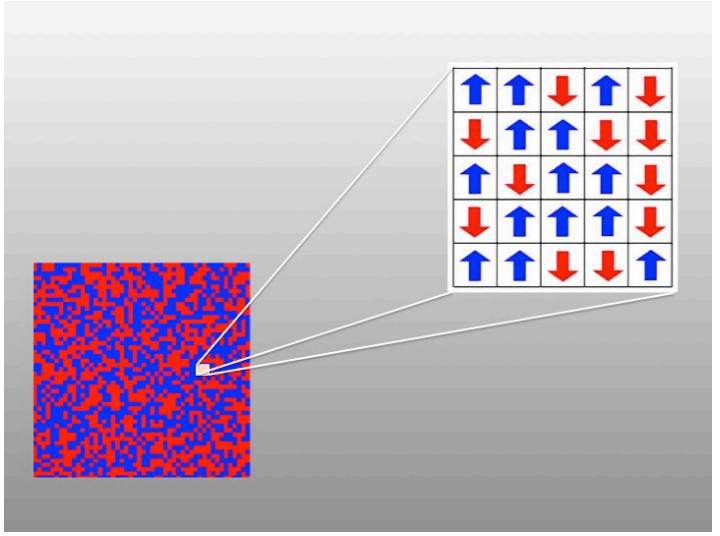


Figure III.2.12: *Ising model.* A two-dimensional Ising system of spins that can only point up or down. Here the system is in a disordered state, where the spins are randomly pointing up or down. If you think of these as nuclear spins, you see that the spins are neatly ordered spatially on a cubic crystal, but that the spin orientations are disordered. So, order and disorder can peacefully coexist if they refer to different degrees of freedom.

nearest neighbors, and the contribution to the energy of any pair of neighbors is,

$$H(\sigma) = -\sum_{ij} J_{ij}\, \sigma_i \sigma_j,$$

where J_{ij} is the interaction parameter. If $J_{ij} = 0$ there is no interaction, whereas if the coupling is constant and positive, $J_{ij} = J > 0$, then we have a ferromagnetic system, and if the constant J is negative we have the anti-ferromagnetic case. If the couplings J_{ij} are chosen randomly, then we speak of a *spin glass*. For simplicity we have left out a term for the coupling of the spins to an external magnetic field. Let us consider the ferromagnetic case, If a pair of neighbors has the same spin, the contribution to the energy is minimal, whereas if the spins are opposite the contribution is maximal. The total energy equals the sum of all pair contributions. For the ferromagnetic case, the minimal energy configuration is therefore

the one where all spins are the same, either all up or all down.

The Ising partition sum. The probability for a configuration to occur is given by the Boltzmann factor we introduced in the section on Statistical Physics in Chapter I.1:

$$P(\sigma) = \frac{e^{-\beta H(\sigma)}}{Z_\beta} ,$$

where the normalization factor Z_β is the *partition sum*:

$$Z(\beta) = \sum_\sigma e^{-\beta H(\sigma)} .$$

Having the probability distribution of configurations, we can define averages, or expectation values. The free energy is defined as $F = -\beta^{-1} \log Z$, and the thermal equilibrium states correspond to the minima of the free energy.

Ising magnetization. To obtain the magnetization we first average the spin over all sites in a given configuration: $M_\sigma = \sum_i \sigma_i / N$, the thermal average is then given by

$$M = \langle M_\sigma \rangle_\beta = \sum_\sigma M_\sigma \, P_\sigma .$$

In Figure III.2.13 we have depicted three configurations, representing the ordered and disordered phases, wich a critical configuration in between.
Order. In the ordered, low temperature phase the domains are macroscopic (the lowest energy configuration is just a single domain with all spins up or all spins down). In the ordered phases the magnetization would be $M \neq 0$.
Disorder. On the right we see a configuration corresponding to the high temperature disordered phase, where there are basically no domains. The individual spins are just randomly pointing up or down, and consequently the magnetization would equal zero.
Critical. In between is the critical case where the temperature equals the critical temperature T_c, where there are domains of all possible sizes. In fact this critical case is special in the sense that it is *scale invariant*, meaning that

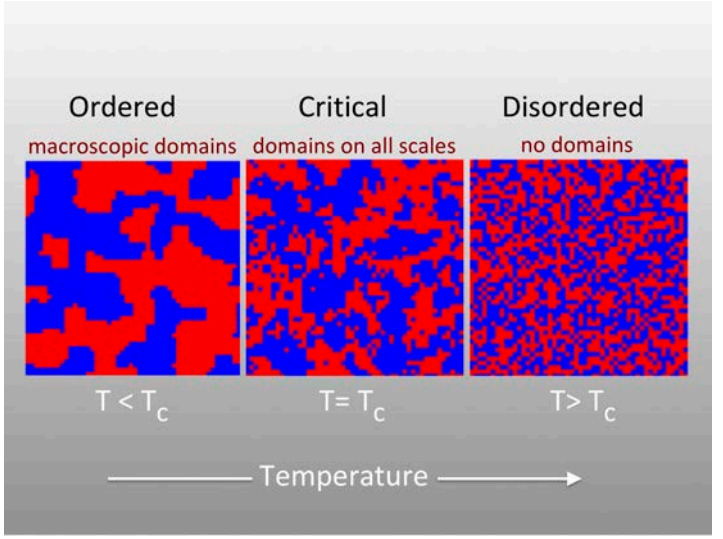


Figure III.2.13: *Magnetic order and disorder.* We see the states of an Ising model without external magnetic field. At low temperatures the state is ordered, and spins are aligned over macroscopic distances, while at high temperatures the state is disordered and there are no domains, just individual spins randomly pointing up or down. In between there is a critical point, where there are domains of all sizes. The critical Ising model is scale invariant.

if you enlarge the picture and cut out a piece of the original size, it would not be possible to distinguish it in a statistical sense from the original one. It is self-similar in a statistical sense.

Mean field theory. One can make an illuminating approximation of the model as a *mean field theory*. One approximates the spins by the local magnetization field $M(x)$. Clearly this approximation will break down for small distances. It is possible to write an effective free energy $F(M, T)$ in terms of this field $M(x)$ this is known as the Landau theory. Because of the symmetries in the model it will only have even powers of the field and in low order it will look like:

$$F(M, T) = \mu M(x)^2 + \lambda M(x)^4 , \qquad \text{(III.2.1)}$$

where the parameter $\lambda > 0$ (the free energy is bounded)

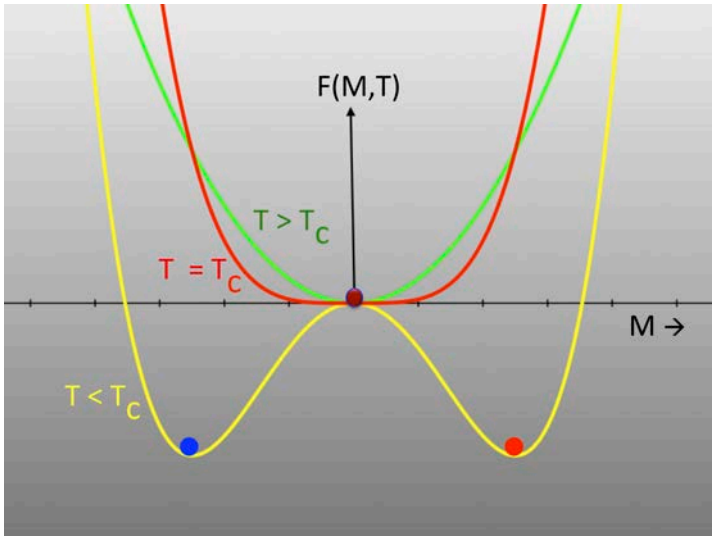


Figure III.2.14: *Second-order transition.* We have plotted the free energy F as a function of the order parameter M (magnetization) for three values of the temperature. At $T > T_c$ the symmetric minimum is at $M = 0$ (no magnetization). At $T < T_c$ the minimum is at $M \neq 0$ (spontaneous magnetization), and the system will 'choose' the red or the blue minimum. This is an example of *spontaneous* symmetry breaking.

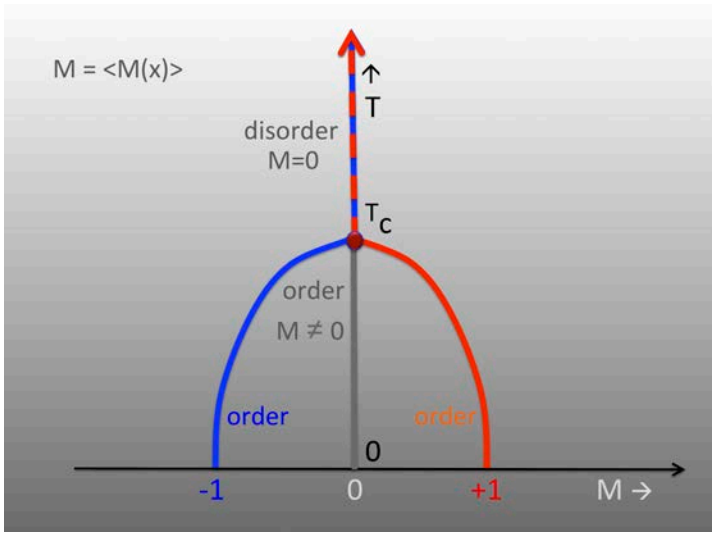


Figure III.2.15: *Ising model phase diagram.* The vertical axis is in fact the phase diagram of the Ising model (without external field). It has only one control parameter which is the temperature. We have plotted the average spontaneous magnetization as a function of temperature. If we lower the temperature the minima of the free energy in the previous figure trace out the blue and red curves giving M for $T < T_C$.

and the other parameter μ has a temperature dependence which near the critical point is given by $\mu = \mu_0(T - T_c)$. In Figure III.2.14 we have plotted the free energy $F(M, T)$ of the system as a function of the average magnetization and the temperature. We see from the figure that the minimum of the free energy for $T > T_c$ yields the value $M = 0$, and for $T < T_c$ we see that the minimum of the free energy corresponds to a non-zero value for M. The latter is the situation where the symmetry of F is spontaneously broken in the sense that the system must choose one of the two degenerate groundstates, with all spins up or all spins down. For $T = T_c$ the system is in the critical state, where the free energy curve flattens out ($\mu = 0$). The vanishing of the quadratic curvature term means that the spin wave excitations have effectively a zero mass (they are 'gapless'). And this is what gives rise to the power law behavior of the correlation functions as we will discuss next.

In Figure III.2.15 we have summarized the results. Along the vertical axis we have a one-dimensional phase diagram with temperature as the only control parameter. For low temperature the phase is ordered, and above the critical temperature it is disordered. In the same graph we have plotted the order parameter, which is the magnetization M along the horizontal axis. The magnetization tends to $M = \pm 1$ as temperature goes to absolute zero. We see that the order parameter as a function of temperature changes continuously in this case, which means that we are dealing with a second-order phase transition.

Correlation functions. A meaningful probe of order and in particular of critical behavior are the spatial correlation functions for large distances. For the Ising model, one calculates the thermal average of the product of two spins σ_i and σ_j but now as a function of their separation $|i - j|$. The

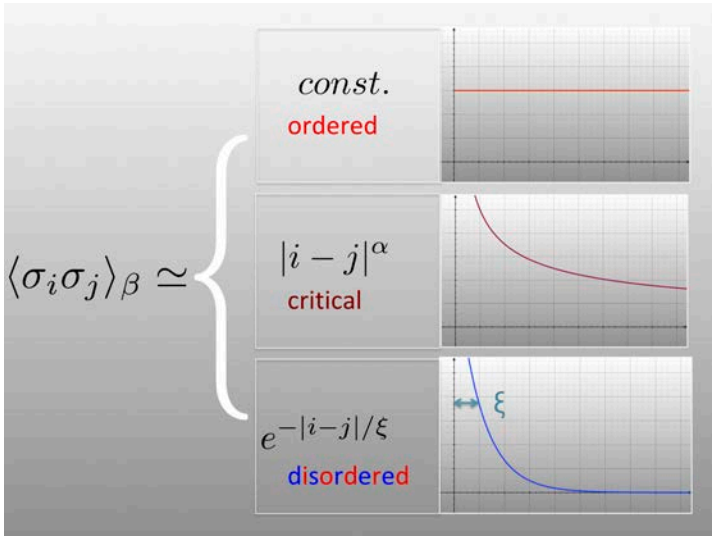


Figure III.2.16: *Correlation functions.* The typical behavior of the spin-spin correlation function $\langle\sigma_i\sigma_j\rangle_\beta$ in the three regimes of the Ising model.

expression is as follows:

$$f(i-j) = \langle\sigma_i\,\sigma_j\rangle_\beta\,. \qquad \text{(III.2.2)}$$

It is simplest to first consider the case at low temperature where there is long range order. This would be reflected in the correlation function to be a non-zero constant. On the other hand, if the system is disordered one expects the correlations to be short range, and indeed the the correlation function can be calculated to decay exponentially over a characteristic length called the *correlation length* ξ. We have summarized the distinct functional behavior of the correlation functions in the three regimes in Figure III.2.16.

Critical behavior. The behavior at the critical point, the phase transition itself, is of great interest. It turns out that the transitions show a high degree of *universality* The correlation functions for example, behave as *power laws*, which means that for large x they behave like $f(x) \simeq x^\alpha$. Such functions are characterized by a power α which is called a *critical exponent*. These exponents express the characteristic quantitative behavior of correlation functions in the critical state, between the ordered and disordered phase. In fact as we approach the critical point from the disordered side on finds that the correlation length $\xi(T)$ diverges, so, $\lim_{T\to T_c}\xi(T) \to \infty$. This is precisely why the exponential decay law in the disordered phase changes to a power law at the critical point.

Universality. It turns out that different types of systems have identical critical behavior meaning that they have the same set of critical exponents at the critical point. These exponents do not depend on the microscopic details of the model but rather on the number of dimensions and the symmetries of the system. The fundamental symmetry underlying second-order phase transitions is *scale* and *conformal invariance*, which can then be extended in various ways to obtain the different universal behaviours. So the critical behaviour of the 2 dimensional Ising model can for example be described on a free massless (Majorana) fermion field. Which means that the spin and energy correlation functions of the two models show exactly the same critical exponents. So, it is also in this field of research that symmetry arguments can greatly advance your understanding observed phenomena. The critical exponents label the representations of the group of certain conformal symmetries in two-dimensions.

Anti-ferromagnetism. Now in magnetism there could be another type of order referred to as anti-ferromagnetism, where the neighboring spins tend to point in opposite directions. This corresponds to choosing the coupling parameter J in the energy expression to equal $J = -1$. The ordered, low temperature, lowest energy configuration now corresponds to a red/blue checkerboard configuration. And the magnetization as defined above would also give zero for this ordered phase. This just illustrates the fact that one has to have some clue or make an educated guess, about what the state looks like before one can come up with a sensible type of order parameter. Here we can the repair

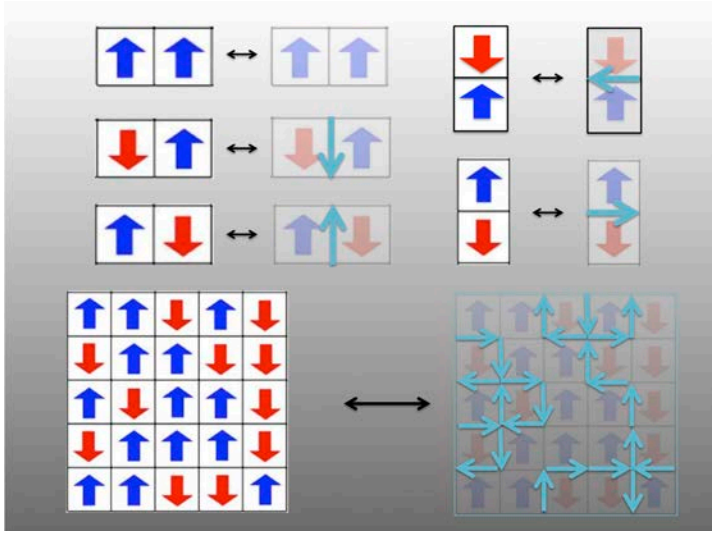


Figure III.2.17: *The loop representation of the Ising model.* We illustrate an equivalent or 'dual' representation of the Ising model where the states are represented as connected oriented paths along the links of the lattice. For any possible pair of neighbors there is a unique prescription. If the paths cross at some vertex there are always two arrows pointing towards and two arrows pointing away from the vertex.

the definition of the magnetization quite simply by adding an extra minus sign on all odd sites, for example. And – as we will see – there are ordered phases in the quantum regime where there no local order parameter can be defined.

Domain walls and defects. If we think again of the energy associated with a neighboring pair, we have $\varepsilon = 0$ if the spins point in the same direction and $\varepsilon = 1$ if they are different. Now with this we can construct a *dual representation* of the Ising model, in terms of oriented contours along the edges of the (dual) lattice. We have depicted this correspondence in Figure III.2.17, For any pair of neighbors we draw an arrow along the edge they have in common if the spins are opposite, or no arrow if the spins are the same. If you now look at a large configuration, then the spin configuration uniquely corresponds to a configuration of oriented lines. There is one subtlety that is clear from the last picture in the figure, if two lines cross, then you always have two arrows pointing in and two out, and this in turn means that there are two options for how to connect the lines at the crossing. If we have a blue domain inside a red domain, that would yield a closed boundary oriented anti-clockwise, and if we exchange the colors, the orientation would flip to clockwise. This representation in terms of these boundary contours or *domain walls* immediately makes manifest where the energy is located. The walls cost energy (because they coincide with a pair of differing neighbors), and the total energy equals the total length of the domain walls. In the ferromagnetic ground state there are no walls, and therefore a domain wall is called a *defect*. It is a *topological defect*, away from the boundaries of the sample the walls form closed loops which cannot break. The loops can grow or shrink, they can join or break up, they can disappear or being created, but a wall cannot have an endpoint in the sample. So, you can also think of the Ising model as a 'gas of loops', with the additional property that the loops don't intersect. You may check this by looking at any would-be intersection of the walls and note that the two ingoing arrows can be connected to the two outgoing arrows only in two ways. Drawing these one finds that they do not cross, indeed. the loops avoid themselves and others.

A dual representation. The two dual representations, one by spin and the other by loop configurations, provide two complementary perspectives on order versus disorder. Starting in the ferromagnetic phase from zero temperature, there are no defects, and it is by raising the temperature that the loops are created, and by the time we are in the disordered state, the loops have 'condensed,' there are defects everywhere. A maximal energy state is one where there is a defect on every link which happens to correspond to a perfectly anti-ferromagnetic state. And indeed changing the sign of the neighbor-coupling J exactly exchanges the highest and lowest energy states of which there are two each.

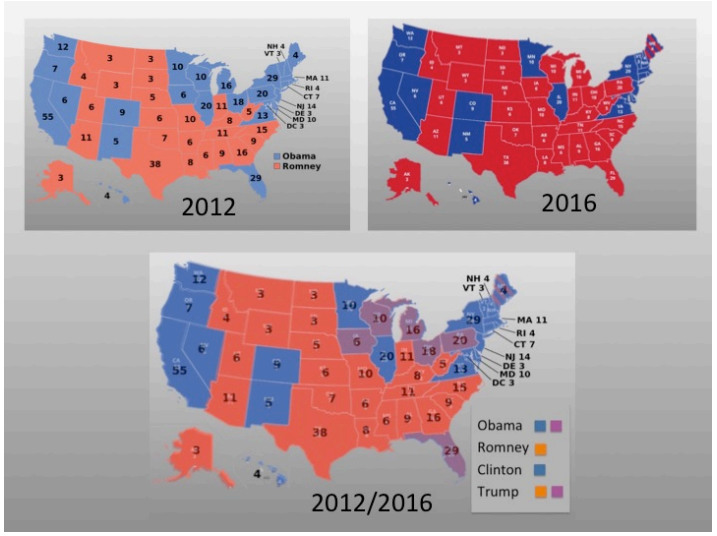


Figure III.2.18: *US voting patterns.* An Ising model representation of voting patters of the 2012 and 2016 elections. In the bottom figure you see the shift. Indeed, the swing states are on the boundary, the shift involves moving the 2012 boundary. If you create a new island, you are considered a defect and it costs a lot of energy, there is a high threshold. Building domain walls is also costly but not as expensive as creating a new domain.

These considerations illustrate a quite general principle, that defining a certain type of order in a system usually also implies the existence of certain types of defects, both topological and non-topological. This is not only true for spin systems but for most forms of order. As will be discussed in the next section, crystals for example have all kinds of defects, of which the *dislocations* and *disclinations* are the most well-known. These defects have their own dynamics, for example if we prepare the spin system starting from high temperatures by *quenching* it, meaning cooling it fast, then the loops will not have time enough to annihilate and the defects get frozen in. If, on the contrary, we cool it slowly, then we may end up with a perfect ground state as the defects had enough time to pair up and annihilate each other.

Swing states

G: Hey Orange! I really like that stuff you are talking about.
O: Thank you Green. It took me quite some effort to master this subject, so I am glad to hear you like it.
G: You know, Orange. I think this stuff may have great applications.
O: But Green, this is pure science just for the sake of ....
G: All that blue and red, that order and disorder, those arrows up and down. It really did make me think of the elections!
O: But Green, ...
G: Those walls, you know. And how hard it is to create blue bubbles in the red domains.
O: But Green, ...
G: You see, if you take the voting patterns of 2012 and you take those of 2016, and you look at what happened.
O: But Green, ...
G: Yes, Orange, yes! Look at that, the swing states are right there bordering on the walls. That's exactly where all their campaign money and energy went, and yes, that's where they got the walls moving. Chr chr.
O: Green! Stop it.
G: And no red bubbles in the blue, and no blue bubbles in the red. Just like you said.
O: That's no science, Green!
G: Hey those swing states are just defects, and nothing happens elsewhere.

O: Stop it!
G: I wouldn't call that a landslide! It's all in the margins, Orange. In spite of all excitement and heated discussions, we are dealing with Ising system at low temperature, with some domain walls frozen in. Don't you think that is a comforting thought.
O: Oh, Green, I wished I never told you.
G: The Ising model of voting! Chr chr. Maybe we should start working on that phase transition, Orange! I mean, what would it be like to live in an anti-ferromagnetic country? They call it disorder, but didn't you just say that it just a different type of order? You know, the colors mix well, and I didn't see a glass ceiling either. There are so many walls, that it is just like having none!
O: Oh no....
G: Oh Yes. I think we should start working on 2020 and 2024 elections including terms for fraud and outcome denial! Chr chr. □

It directly follows from simple energy considerations that it costs more to create a red site in the middle of a blue domain (four units of energy), while moving a red boundary, which means changing a blue to red site at a boundary always costs less. From this local energy perspective it is also clear that domain walls will have the tendency to straighten out.

Defect condensation and dual order. The state we have described as disordered, where the spins are randomly distributed, can be considered from the dual point of view as a state where there are defects all over the place. If we were to define a dual order parameter measuring the average number density of wall segments or links on the dual lattice, it would be non-zero. In other words, it is a kind of dually ordered phase where the defects have condensed.

Crystal lattices

Symmetry reigns. At low temperatures or high pressure, atoms (or ions) tend to settle down in periodic arrays which correspond to a crystal lattice. A characteristic of such a lattice is that it is periodic, and there is a certain basic geometric pattern – called a *unit cell* – that repeats itself over and over again. So if you move the (infinite) lattice over a certain distance in certain directions it looks exactly the same, and the same is true if one rotates around certain axes by particular angles or reflects the lattice about in certain planes. The lattice can be characterized by the set of symmetry operations that leave the lattice invariant. These operations form intricate infinite discrete groups, consisting of discrete translations and rotations.

Wallpaper groups. The five basic space filling lattices in two dimensions and their corresponding space groups have been constructed, they form the so-called *wallpaper groups* and there is a total of seventeen of them.

The Bravais lattices. The space-filling crystal lattices have been classified by the nineteenth century French mathematician Auguste Bravais. In two dimensions there are five different lattices. In three dimensions there are seven basic lattices to which special points may be added, making a total of 14 Bravais lattices. Not surprisingly there is an awesome jargon that comes with them in order to distinguish them, involving terms like *cubic-face-centered, orthorhombic, triclinic, rhombohedral* and so on. In particular *cubic-face-centered* sounds to me like a fancy AI surveillance algorithm!

For the 14 space-filling, three-dimensional lattices, the space groups have been fully classified and everything is known about all 230 of them. This means that also the point groups preserving the unit cell in three dimensions are known and there are 32 of them.

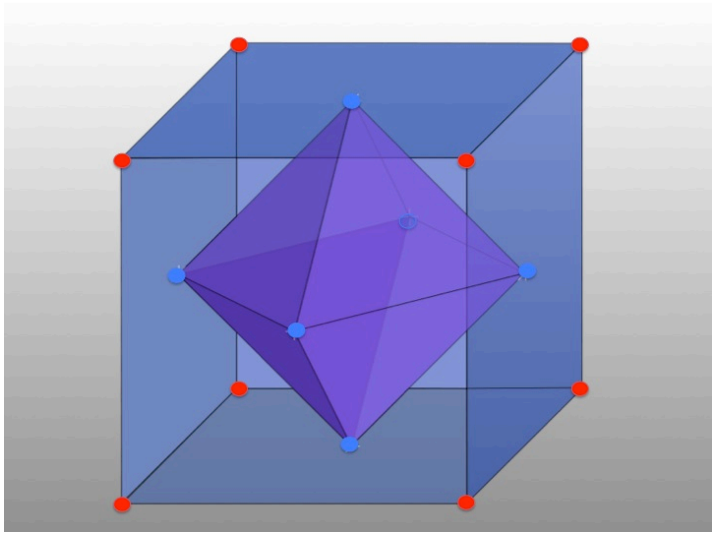


Figure III.2.19: *Symmetries of octahedron.* The embedded octahedron has the same symmetry group as the cube.

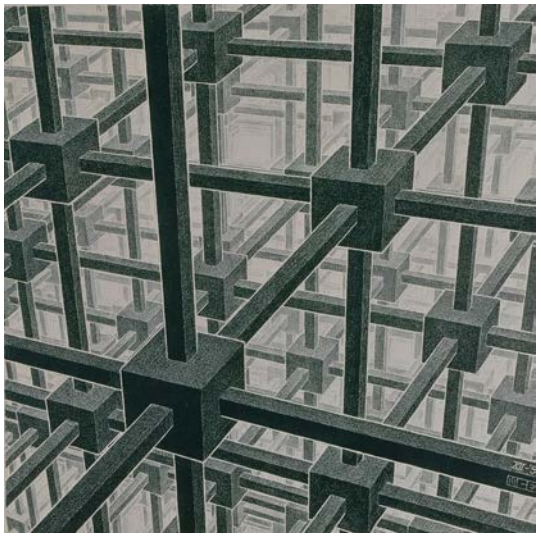
X-ray diffraction and more. Crystal lattices can be studied experimentally by short wavelength photons (X-rays). The X-rays scatter from the nuclei on the lattice sites and the scattered waves will interfere with one another. So, whether we get reflection of diffraction depends on whether the interference of the many scattered waves is constructive or destructive. The crystal has planes of atoms in various directions and the photons may be diffracted or reflected depending on whether their momentum satisfies certain conditions which are determined by the specific geometrical properties of the lattice. From the reflection and diffraction patterns one can then reconstruct the geometry of the lattice.

This widely applied technique of studying molecular order whether it is lattices or complicated molecular structures like DNA[3] was invented by the British physicists William Henry Bragg and his son William Lawrence who shared the Nobel prize for Physics in 1915. The application of the technique to the complicated molecules of life was pioneered by Max Perutz, an Austrian refugee, who got a position at the Cavendish laboratory in Cambridge with the Braggs. Nowadays we can probe the surface of solids on atomic scales by advanced microscopes, the *scanning tunneling microscope* (STM) or the *atomic field microscope* (AFM). But the 3-D imaging is still of the diffractive type. These probing techniques are – not surprisingly – based on quantum principles themselves.

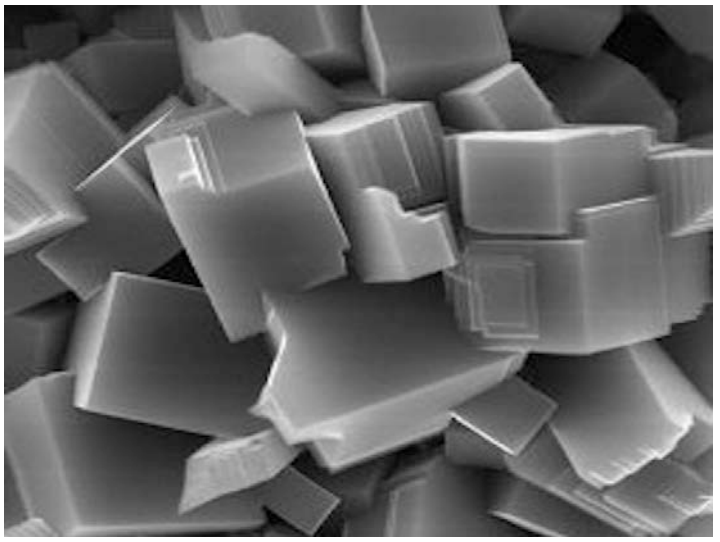
There is the remarkable fact that if you want to probe nature at some scale then nature often also provides you with the tools which are operative at the same scale, that allow you to build suitable probing devices. It is a matter of giving and taking. This is true for atoms with visible light, for nuclear structure using nuclei (alpha particles), and is true for genetic manipulation using all sorts of enzymes etc.

Kitchen salt or the cube. Let us now look in more detail at some three-dimensional lattices. A well-known example in three dimensions is the kitchen salt or sodium chloride (NaCl) crystal, which is a simple cubic lattice with the sodium and chloride atoms occupying alternating sites (see Figure III.2.20(c)). The *point group* of the cubic lattice, which is the symmetry group of the cube, is surprisingly rich and consists of 24 elements. As indicated in Figure III.2.20(d), it has four threefold axes (rotations around main diagonals), three fourfold axes (around lines through centers of opposite faces), and six twofold axis (through centers of opposite edges). This group is denoted by O and called the octahedral group, because it is also the symmetry group of the octahedron obtained by drawing the planes through the face centers of the cube, as one may see from Figure III.2.19. Indeed, correcting for the identity element we verify that the group has indeed $1 + 3 \times 3 + 4 \times 2 + 6 \times 1 = 24$ elements. The transformations we discussed so far are all rotations, but there is one more transformation that leaves the cube invariant,

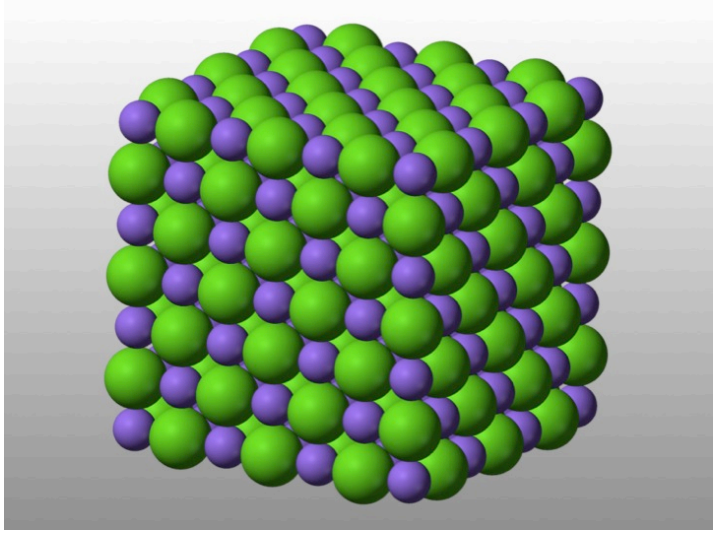
[3] There is the (in)famous story that Francis Crick and James D. Watson discovered the structure of DNA in 1953 after Maurice Wilkins had shown them a diffraction pattern measured by Rosalind Franklin at King's College London. It held the clue to the spatial structure of the double helix.



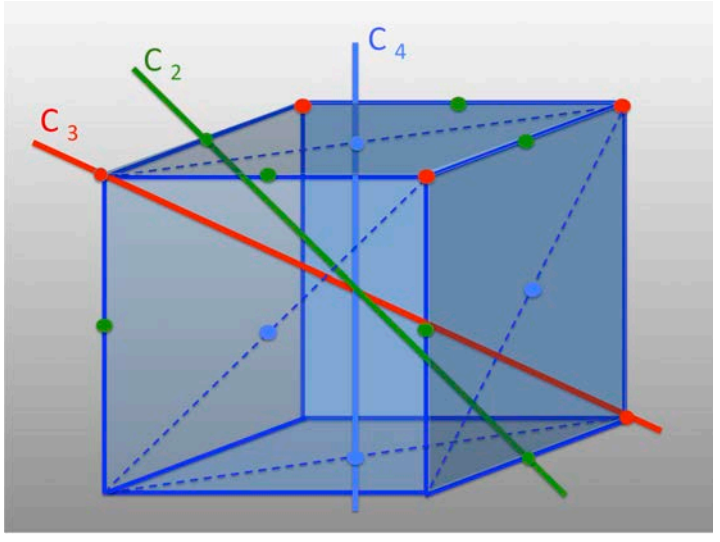
(a) The *Cubic Space Division* by M. Escher. It is invariant under translations by the lattice constant a along the x, y and z axes. (© 2023 The M.C. Escher Company.)



(b) Kitchen salt crystals of about 10 micrometers. Image taken with environmental scanning electron microscope (ESEM) at 950^{o} C .



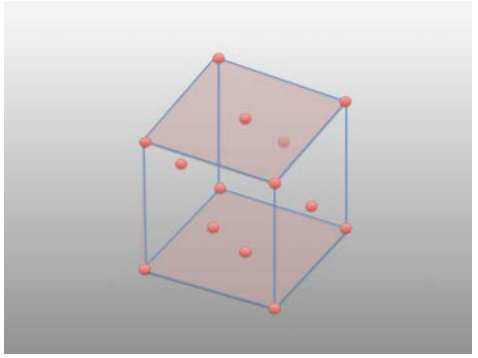
(c) The crystal of kitchen salt or sodium chloride ($NaCL$). It is a simple cubic lattice with alternating sodium (purple) and chloride (green) ions. (Source: MIF Univ. of Calgary.)



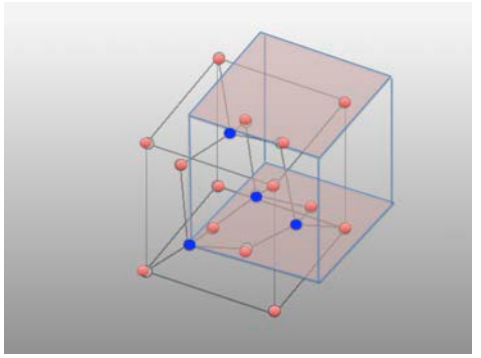
(d) The symmetries of a cube. It has three fourfold axes (blue), four threefold axes (red) and six twofold axes (green). The set of all transformations that leave the cube invariant is the *orthohedral group* O ; it has 24 elements.

Figure III.2.20: *The symmetries of the cube*.

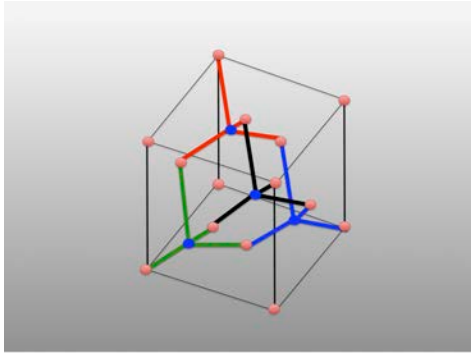
(a) The facets of a diamond are designed to maximize its reflections.



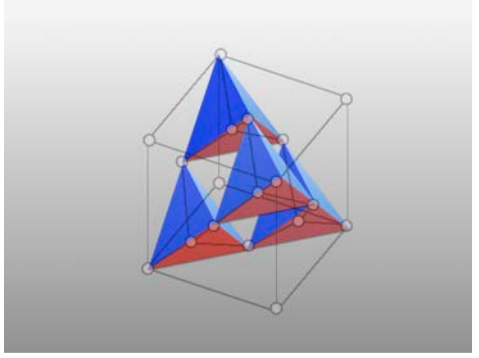
(b) A cubic face-centered (fcc) lattice cell.



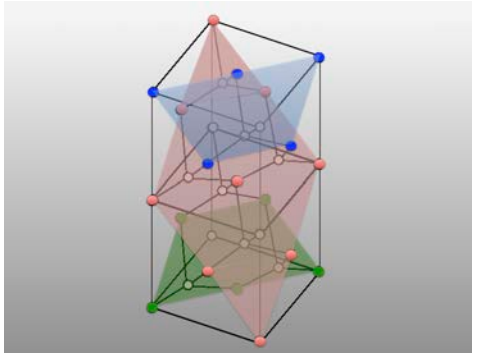
(c) A second fcc lattice superimposed at the point $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ in blue.



(d) The resulting diamond lattice as a stacking of tetrahedra.



(e) The lattice as a stacking of planes with tetrahedra (with center).



(f) The lattice stacking of planar triangular lattices.

Figure III.2.21: *The diamond lattice.* The intricate diamond lattice and some ways to look at it which display different aspects of its symmetry.

namely inverting all coordinates, which amounts to mirroring every point of the cube in the origin. This is called the *inversion* or *parity* operation P. If we add this transformation, we get a group denoted by $\mathrm{O_h}$ with 48 elements. This group is non-abelian (not all elements commute with each other) as one can easily check.

Diamond. Crystal lattices clearly exhibit intricate and aesthetically pleasing features and coincidences. A nice example is the diamond lattice which is more involved, and we explore different perspectives on it in Figure III.2.21. The structure is built-up of two cubic face-centered (fcc) lattices shifted with respect to one another (III.2.21(b) and III.2.21(c)). The result is a perfect three-dimensional stacking of tetrahedra (III.2.21(d) and III.2.21(e)): the corners are all on the first fcc lattice while the centers are on the second fcc lattice. The lattice can therefore also be viewed as a stacking of planes with tetrahedra. One can go one step further and think of the whole lattice as a stacking of pairs of strictly identical triangular lattices, one of each fcc lattice. In Figure III.2.21(f) we show the three top layers of subsequent pairs, which all belong to the first fcc lattice. Projecting all the points down along the body diagonal, which is perpendicular to the layers one finds that there are three inequivalent triangular lattices in the figure corresponding to the blue, red and green layers.

The uses of symmetry. It turns out that the symmetry group tells us a lot about the physics of the system; it not only characterizes the stable equilibrium or ground state but also yields a natural labelling of the low energy modes that can propagate through the system. The symmetry teaches us also about properties of the spectrum of electrons. And finally, the symmetry group of the lattice determines the possible lattice imperfections or *defects* that may occur.

As we live in three-dimensional space most of us will agree that our analysis should stop there. The classification of space groups and lattices in higher dimensions is to be considered a mathematical pastime at best. But nature had a surprise in store. Who would have expected that higher-dimensional regular lattices would rear their heads also in our three-dimensional world in the guise of so-called *quasicrystals*.

This provides another striking example of the 'unreasonable effectiveness of mathematics in the natural sciences,' which refers to the title of a famous lecture by Eugene Wigner who got the Physics Nobel prize exactly for his work on group theory and its many applications in quantum theory. We will return to quasicrystals towards the end of this section.

Crystalization and symmetry breaking

We introduced and expanded on the concept of symmetry breaking in Chapter II.6. It has many beautiful applications in condensed matter, and in particular also in the theory of crystallography. In this section we explore two representative examples.

The concept. Suppose one of the atoms in a simple cubic lattice is of a different type, say it is has a different color, then we may ask for the transformations that leave not only the cube invariant but also keep the colored atom in place. For this case the answer is quite obvious from Figure III.2.20(d). If we ask which transformations leave not only the center but also one of the red dots in place, then we are only left with a single threefold axis. This means that the rotation group $\mathrm{G} = \mathrm{O}$ is reduced to, or as is often said, *broken* to, $\mathrm{H} = \mathrm{C}_3$. This reduction of the symmetry from a group G to the so-called *residual symmetry group* H, which is a subgroup G, means that certain degeneracies in the spectrum that occurred in the unbroken situation will now be lifted. So one could say that breaking the symmetry allows for less uniformity and more differentiation.

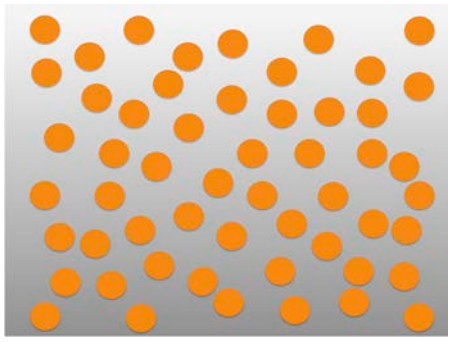
Symmetry breaking is therefore an invaluable tool to analyse and interpret experimental data. In particular if we have certain external parameters we can change, like temperature or electric or magnetic fields, it may be that what appeared as one state breaks up in a set of different states. Then different states with the same energy may split up in states with different energies, much in the way we discussed in Chapter I.4 in relation to the Zeeman effect. There the spherical symmetry of the atom was broken by the direction of the external magnetic field, and the spectral line was split because the degeneracy of the states was lifted. Symmetry breaking may also happen spontaneously if one lowers the temperature, as is the case with 'spontaneous magnetization' in a magnet as described by the Ising model. Even without an external magnetic field the spins may line up because of their local ferrromagnetic interactions.

You now may also understand that the formation of a lattice itself, the process of crystallization, is an example of spontaneous symmetry breaking. You should think of starting with a liquid which we envisage as a continuum. If you are at some point in the liquid it looks the same, independently of what point you chose, and it looks also the same in all directions. A simple fluid is therefore said to be *homogeneous* and *isotropic*. This translates in the statement that the symmetry of a simple liquid consist of all rotations by any amount about any axis, and also of translations in any direction by any amount. Clearly this group is continuous and is called the *Euclidean group* E_3 of three-dimensional rotations and translations we mentioned before. It is the symmetry group of empty three-dimensional Euclidean space. So crystallization is a process where the symmetry gets broken from the Euclidean group to the symmetry group of the lattice, which is a discrete subgroup of E_3.

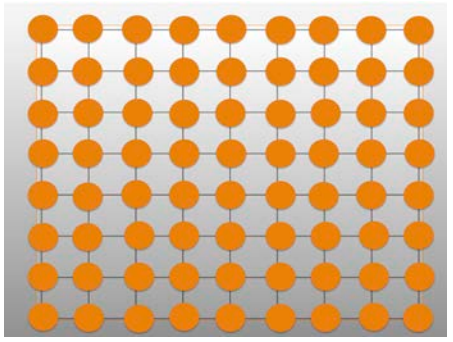
Goldstone modes. We have in the section on symmetry breaking of Chapter II.6 mentioned how breaking of a continuous (global) symmetry leads to the existence of massless modes. This is precisely what happens upon crystallization, where the Euclidean group gets broken to the discrete lattice group. The low energy modes correspond to the sound modes that can propagate through the crystal. They are the Goldstone modes which are associated with the breaking of the continuous translational symmetries of the perfect fluid. In Figure III.2.22 we give the pictorial account. From (a) to (b) the crystallization takes place. In Figure (c) we have sketched a sound mode corresponding to a longitudinal pressure or density wave that propagates through the crystal.

Topological defects. There is an additional observable consequence of broken symmetry in the situation we are discussing. Broken symmetries manifest themselves not only in lifting degeneracies and the presence of particular low energy modes, but also in the presence of *defects*, called lattice defects in the case at hand. The theory predicts that if we break the continuous group E_3 to the discrete group of the cubic lattice, we have line defects that we in principle can label by the elements of the symmetry group of the lattice. In a crystal we typically distinguish two kinds: translational defects called *dislocations* and rotational defects called *disclinations*. We have illustrated them for a two-dimensional lattice in the pictures (d) and (e) of Figure III.2.22.

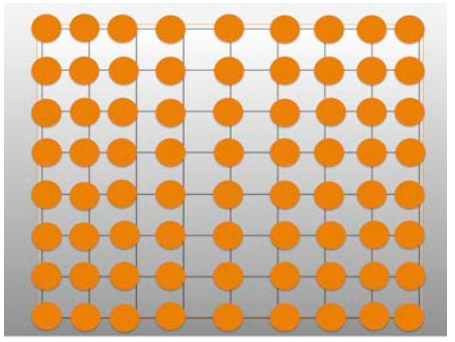
Dislocation. In the bottom left Figure III.2.22(d) the dark atom is special, since it marks the endpoint of an extra vertical layer that does not go all the way up. Note that far away from the marked atom the lattice has restored itself to its normal unperturbed form. The marked atom is an irregularity, a defect. How do you quantify the defect? In this case you should compare the near environment of a normally positioned atom with that of a defect. If you walk around a normal atom like the one marked in the upper right corner, following the blue arrows you see that it takes 8 steps to get back. If you take 8 steps around the defect site, you go one step too far, and you have to move back by one lattice vector (marked in yellow). This translational



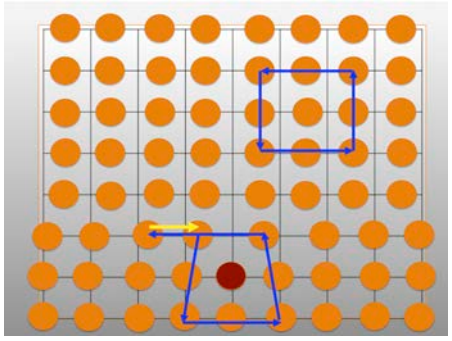
(a) A gas or liquid made of simple atoms. It has no long range order and therefore effectively a continuous translational and rotational symmetry.



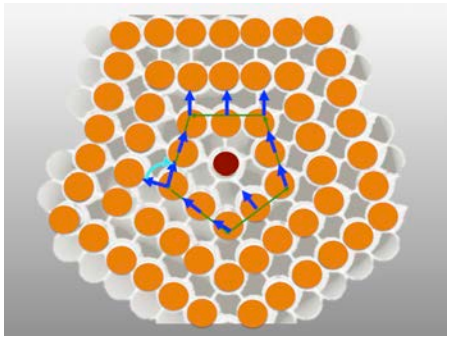
(b) Upon cooling the atoms may 'freeze' and form a regular lattice. The symmetry corresponds to a *space group* consisting of discrete translations and rotations.



(c) A sound wave propagating horizontally. Sound is a periodic density fluctuation in the direction of the motion (longitudinal). The atoms are coherently moved out of their equilibrium position.



(d) The empty site is a *translational defect*, also called a *dislocation*, because when going around it in 8 steps one's position is shifted by one lattice distance. As indicated, away from the defect the lattice is restored.



(e) A *rotational defect* (disclination) related to a rotation over an angle of 90^o . One sees the defect angle if one carries a little vector tied to the local lattice frame around the defect. Starting on the left we obtain a defect angle of 90^o .

Figure III.2.22: *Defects and broken symmetry.* We show two types of lattice defects in a simple two-dimensional crystal.

defect is thus labeled by a translation vector (also called a Burger's vector), and that elementary translation corresponds to a basic element of the discrete translation part of the lattice group. One encounters this one step dislocation on *any* loop around the defect location. Therefore, the defect is uniquely labeled by this group element.

It is easy to imagine that in the process of crystallization such dislocations may form spontaneously. The number of dislocations one finds will depend on how fast we cool the system down. It is clear from the picture that the defect locally deforms the lattice and therefore will carry a certain amount of extra energy. The dislocation in two dimensions is a point defect, and it is stable for a topological reason. You may be able to move it around, but you cannot smoothen it out locally. You can think of the dislocation to be connected (via the extra layer) to either the boundary of the sample or to an 'anti-defect', which means that these defects are locally stable but can annihilate with an anti-defect.

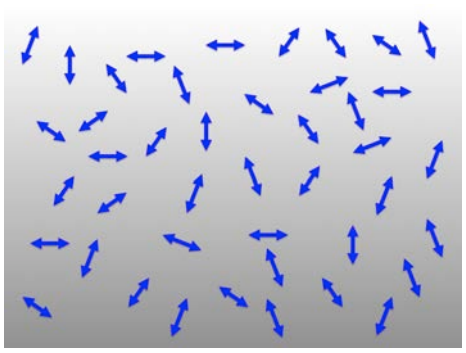
Disclination. In the bottom right picture, we show a disclination or rotational defect. This defect is labeled by the 'defect angle' you encounter as you parallel transport a local lattice vector (or frame) around the defect. If in the figure we take the local blue vector smoothly along the green path around the defect and return to the starting position, the vector has rotated over an angle of 90^o, and again this is an element of the symmetry group of the lattice. This analysis reminds us of our considerations in the section on curved spaces in Chapter I.2, where we discussed this characteristic and called it a non-trivial *holonomy*. It requires a lot of energy to make a disinclination. They may spontaneously form in small samples, and alternatively you can also imagine 'growing' the crystal starting from the impurity outward. That way the fivefold symmetry would be introduced 'by hand.' It is not a lattice in the normal sense because the translational symmetry is broken right from the start of the growing process, that is the price for having a fivefold rotation symmetry in the plane.

Liquid crystals

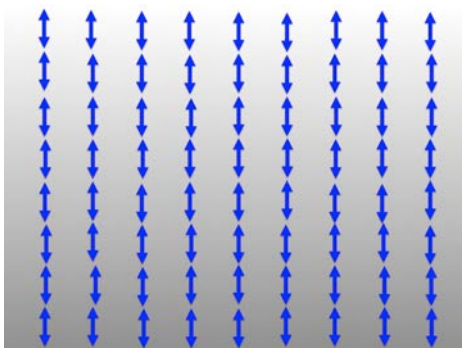
We have alluded to the importance of the shapes of constituent particles for understanding their collective behavior. This hidden underlying geometry is one of the keys to the diversity that is displayed in properties of materials. In the previous section we showed that these shapes can often be translated into symmetries or their breaking. A splendid example of this are the types of order/disorder that arise in soft condensed matter physics, in particular the subject of *liquid crystals* and *nematics*. With the language of symmetry at hand we can give some qualitative characteristics of the materials straightforwardly. The examples are quite easy to visualize and are used to further illuminate the rather abstract notion of symmetry breaking.

Partial order. As mentioned, an ordinary lattice is an example where we break the continuous Euclidean group E_3 down to an infinite discrete group of translations and rotations. It is not so hard to imagine that media can have strange mixtures of order and disorder which are in between a liquid and a crystal. In such cases the translational symmetry is not broken but the rotational symmetry is: the system is partially ordered. These types of systems can easily be visualized by assuming that the building blocks have simple geometric properties, for example they are like tiny rods or pancakes or tetrahedra.

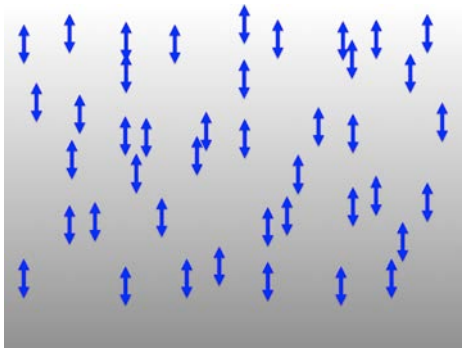
Nematics and smectics. In Figure III.2.23 we illustrate various possibilities if the constituents are rod-shaped. They can form an ordinary liquid or a fully ordered crystal, with both translational and rotational order. In Figure III.2.23(c), however, they form a two- or three-dimensional structure which preserves orientational order with translational symmetry, which is a liquid crystal called a *nematic*. The next picture shows another realization: the rods are oriented along the z direction. Furthermore, the rods form strict horizontal layers, but within the layers there is free motion.



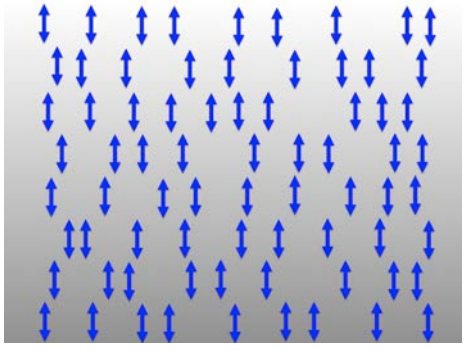
(a) A nematic liquid made of simple rod-shaped atoms. It has no long range order and therefore effectively a continuous translational and rotational symmetry.



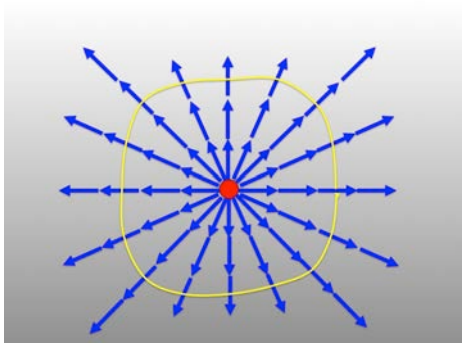
(b) Upon cooling the atoms may 'freeze' and form a regular lattice. Translational and rotational symmetries are broken to a discrete space group consisting of discrete translations and 180^o rotations only.



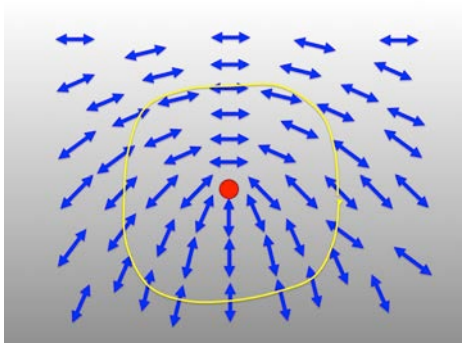
(c) A liquid crystal in which there is still complete translational symmetry, but the rotational symmetry is broken. Such a phase is called *nematic*. There is no positional order but there is orientational order.



(d) This system is called a *smectic*. It is anisotropic, as it is made up of independent layers in which the horizontal translational symmetry is still manifest, but in the vertical direction it is ordered. There is complete orientational order.



(e) A rotational defect (a *vortex*) as it exists in an ordered spin system (represented by ordinary arrows), related to a rotation over an angle of 360^o . This is observed if one follows the direction of the spin vector if one moves around the defect.



(f) This is a rotational defect in a nematic of rods. It is called a *half-vortex* as it corresponds to a defect angle of 180^o . This defect is not possible in a spin system like in (e), going around the direction of the spin arrow would point in the opposite direction.

Figure III.2.23: *Nematics.* Various types of two-dimensional order in a *nematic* system made up of rod-shaped molecules.

This structure is called a *smectic*. A third possibility (not depicted) is called a *uniaxial nematic*, where the rods are vertically stacked in thin filaments. In the direction of the filaments there is the translational order of stacking, but there is no horizontal order across the filaments.

Defects. In Figures III.2.23(e) and III.2.23(f) we have depicted two rotational defects, the first one is an ordinary point defect one may encounter in a two-dimensional spin system or vector field, but of course it can also exist in a nematic. The signature of the defect is that parallel transporting a vector along a closed loop around the defect the spin rotates over 360^o as indicated in the figure. The last picture shows a 'half-vortex', and we see that the configuration is smooth although the rod rotates only over 180^o when taken around. So, it is a point-like defect. This configuration will not form in a spin system because there would necessarily be a discontinuity along a line starting from the defect and ending at the boundary. Such a line would cost much energy and that suppresses the formation. One way to look at this is to say that the half- vortices are 'confined' in the ordered two-dimensional spin system. Indeed, if one cools a spin or nematic liquid rapidly through the transition one usually finds many of the allowed point defects in the (partially) ordered system.

We have illustrated the idea of liquid crystals with a very simple example, but it should be clear that there is an unlimited arsenal of variations and alternatives that has very actively been pursued for example under the name *polymer physics*. As we mentioned Pierre-Gilles de Gennes of the College de France made many invaluable contributions to the early exploration and further development of this field of research.

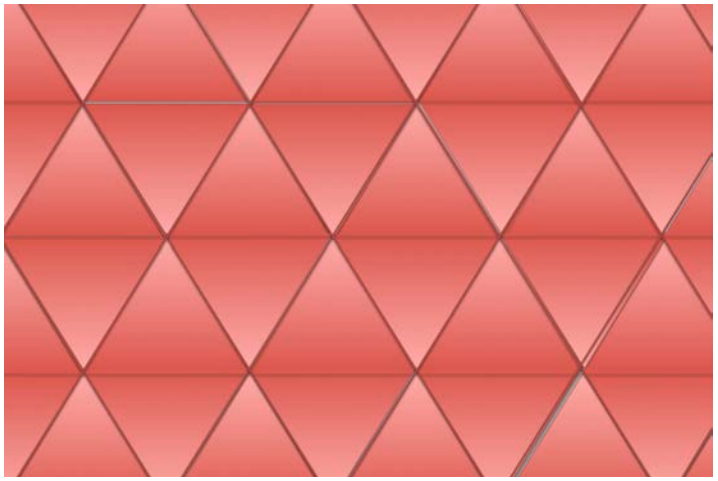
Quasicrystals

Tilings of the plane. In Figure III.2.24 we have depicted some tilings of the plane by simple regular polygons.[4] It works perfectly for triangles, squares and hexagons, but with pentagons (Figure III.2.24(c)) it doesn't quite fit and one cannot tile the plane. A consequence of this is that in the diffraction patterns there can be no signature of a fivefold symmetry. In three dimensions something similar happens, since it is not possible to fill space by stacking dodecahedra which do have fivefold symmetries. The Bravais lattices we discussed before do not admit? fivefold axes and therefore the diffraction patterns of periodic crystals can only have two-, three-, four-, and sixfold symmetries and not have a fivefold symmetry.

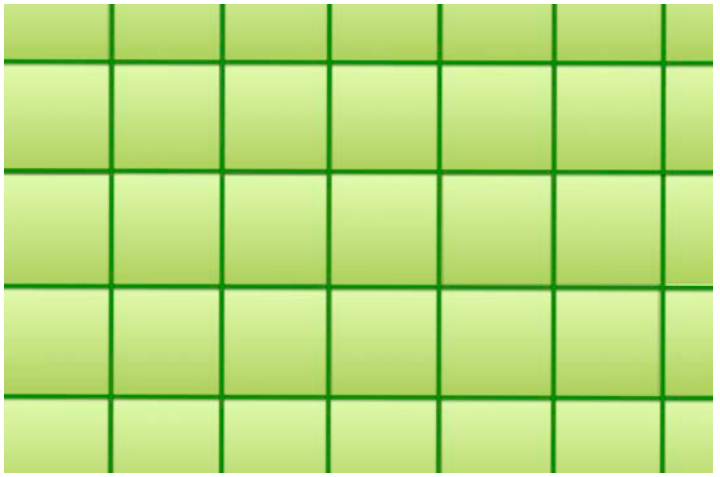
It was a big surprise therefore, when in 1982 the Israeli physicist Daniel Shechtman actually observed a clear diffraction pattern that appeared to come from a perfect crystal but nevertheless showed a manifest fivefold symmetry, like the pattern displayed in Figure III.2.26(c). How could that be? Could there be a nice Bragg diffraction pattern coming from some non-periodic structure? Yes indeed, it turned out that a nice but not perfect diffraction pattern could be generated not only by a perfectly periodic, but also by a non-periodic structure. The system of Shechtman was clearly perfectly ordered, otherwise there would not be such a clear diffraction pattern, but could not be periodic, because that is incompatible with the fivefold symmetry. With his observations the new field of quasi(periodic)-crystals was born. Shechtman received the Nobel prize in Chemistry in 2011 for his remarkable discovery which caused a paradigm shift in the well-established field of crystallography.

Non-periodic tilings. An instance of a quasi-periodic structure is a tiling of the plane by two types of *rombhi*, de-

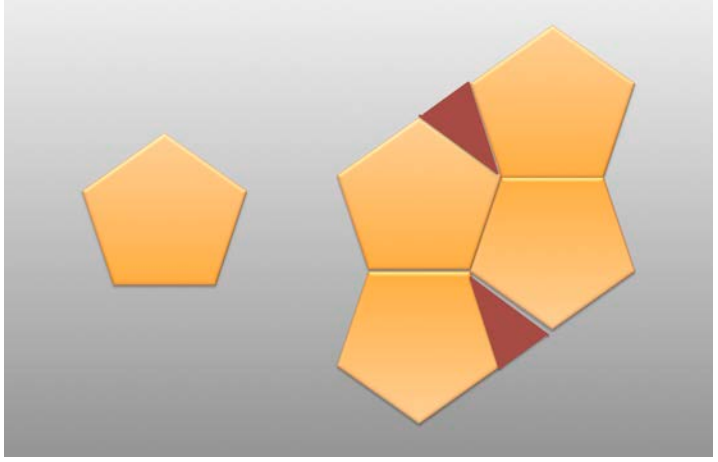
[4] A regular polygon has equal angles and is equilateral.



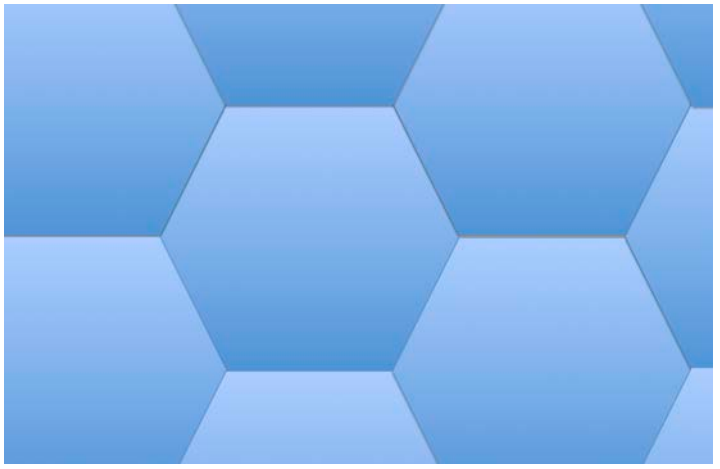
(a) A triangular tiling of the plane.



(b) A square tiling of the plane.

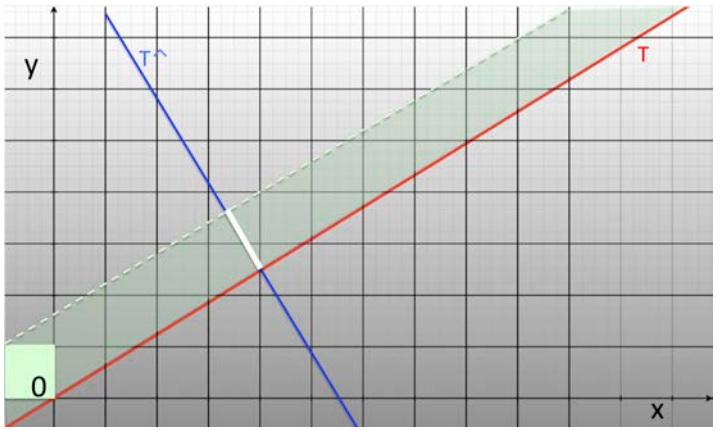


(c) The plane cannot be filled with pentagons.

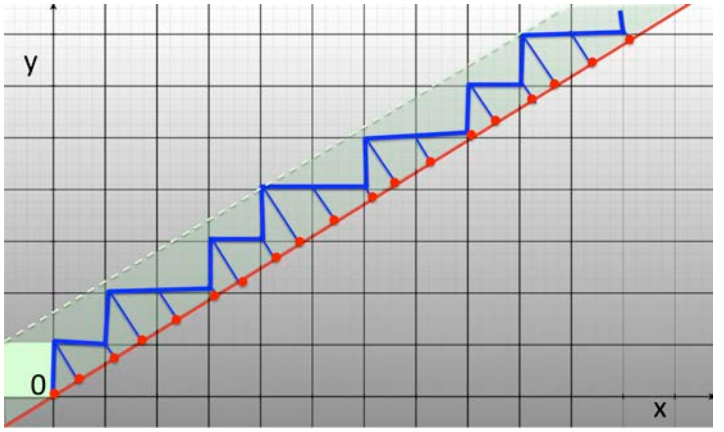


(d) A hexagonal tiling of the plane. Adding the centers would make it a triangular lattice like (a) again.

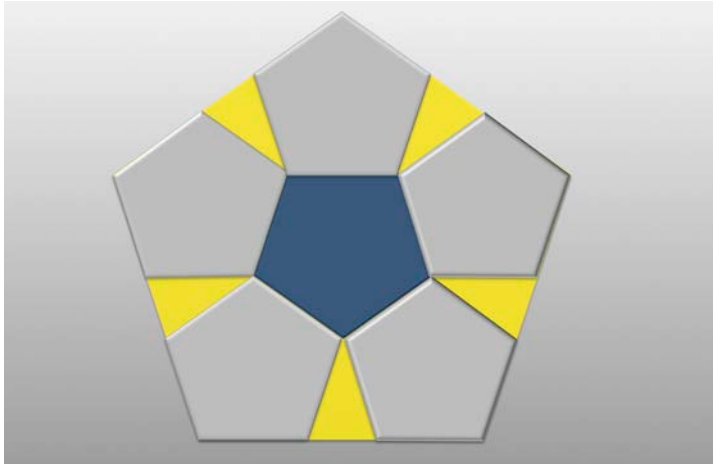
Figure III.2.24: *Polygon tilings.* Possible and impossible polygon tilings of the plane. The regular tilings have discrete translational and rotational symmetries plus reflection symmetries in certain planes.



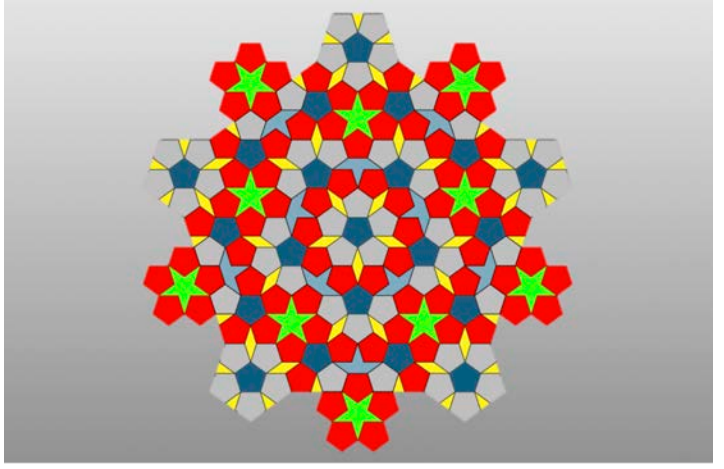
(a) The projection method to obtain a non-periodic tilings. A strip is constructed shifting the basic cell along the subspace T on which one wants to project, and the white segment (subspace) is the intersection of the strip with $T^{\perp}$.



(b) All lattice points in the strip are projected on T producing a non-periodic 'tiling' of the red line (space).



(c) A pentagon filled with six smaller pentagons. Embedding six pentagons into a larger one can be repeated indefinitely, to generate the Penrose tiling P1.



(d) The Penrose tiling P1. The tiling is *self-similar* and basically a *fractal*. Translational invariance has been given up in favor of scale invariance.

Figure III.2.25: *Non-periodic tilings.* Non-periodic but scale invariant tilings of the line and of the plane.

picted in Figure III.2.26(a). This tiling has an approximate fivefold local symmetry. The mathematics of these non-periodic tilings has been developed by the British mathematical physicist Roger Penrose in the early 1970's.[5] They are remarkable in that there is no translation which leaves the tiling invariant. They can have reflection symmetry and for example a fivefold rotation symmetry. But the Penrose tilings have another more subtle so-called scaling symmetry, which means that from any point in the tiling you can blow up or shrink the tiling by a certain amount and it will fit again. This means that such patterns are *self-similar*: they repeat themselves on larger and larger scales and are therefore a special kind of so-called *fractals*.

A one-dimensional Fibonacci tiling. One way to obtain quasicrystals or quasi-periodic tilings is by projecting regular periodic lattices from higher dimensions. We have illustrated this in Figure III.2.25. The top two pictures illustrate the method of going from a simple two-dimensional square lattice to a non-periodic one-dimensional 'lattice'. One first defines the 'physical' one-dimensional space T like the red line in the figures. In this example the line has a slope $2/(1+\sqrt{5})$, which is equal to the inverse of the Golden Mean. This slope is an irrational number which ensures that it will never go through a point of the lattice and that guarantees that the sequence is not periodic. The following step is to shift the two-dimensional unit cell along T, and this defines the light shaded strip along T. Next one projects all lattice points in the strip parallel to the orthogonal subspace $\mathrm{T}^{\perp}$ on T and one gets a non-periodic covering of the line by line segments of only two distinct lengths, being the two different one-dimensional tile types. The sequence of short (s) and long (l) segments forms a so-called *Fibonacci chain*: sl, sll, slsll, $\mathrm{sllslsll}$, ... Each next entry of the sequence is obtained by joining the previous two, which makes the sequence as a whole 'self similar'. Every finite sequence is repeated an infinite number of times, but that does not imply that the chain is periodic. There is also an alternative way to construct the sequence through some 'growing' algorithm. This is a general method that can be used to generate any Penrose tiling and is referred to as the *substitution* or *inflation method*. This is beyond the scope of this book, and we will not discuss it in any more detail.

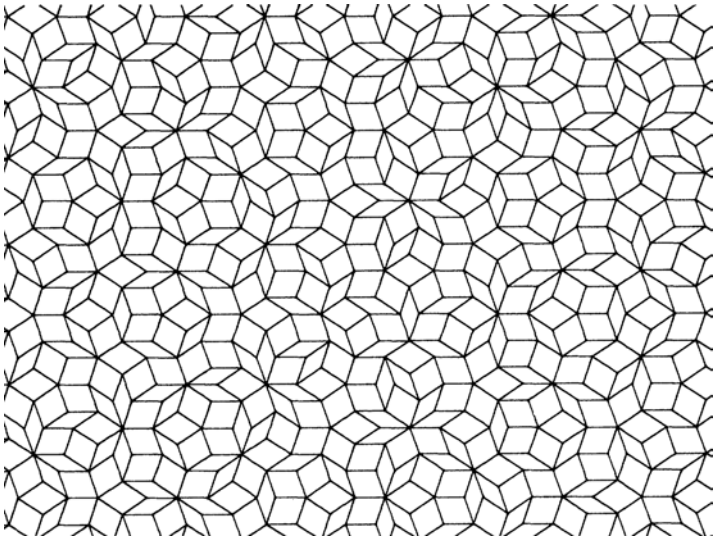
The two-dimensional Penrose tiling P1. Let me now give you an idea how one can obtain a non-periodic tiling in two dimensions with a fivefold symmetry by the projection method. We start with a five-dimensional simple cubic lattice. This lattice evidently has a fivefold symmetry rotating about the diagonal of the hypercube, where the corners on the five coordinate axes are rotated into each other. This is just like the threefold axes of the three-dimensional cube depicted in Figure III.2.20(d). We choose the physical space as a plane that is orthogonal to the fivefold axis. We then move the hypercube over the plane to obtain a five-dimensional layer. All lattice points and edges in that layer can now be projected orthogonally on the two-dimensional physical space, and then a tiling like the Penrose tiling $\mathrm{P1}$ of Figure III.2.25(d) results. The figure shows that $\mathrm{P1}$ needs four types of tiles to fill the plane: the 'pentagon', the 'star', the 'boat' (half star) and the 'lozenge'. The tiling has an approximate 'local' fivefold rotational symmetry.

The projection method allows us to generate all the two- and three-dimensional Penrose tilings. From the figure one may correctly guess that also the $\mathrm{P1}$ tiling also can be constructed from a concentric 'growing' algorithm. Not surprisingly the topic of quasicrystals has given rise to a prolific mathematical literature.

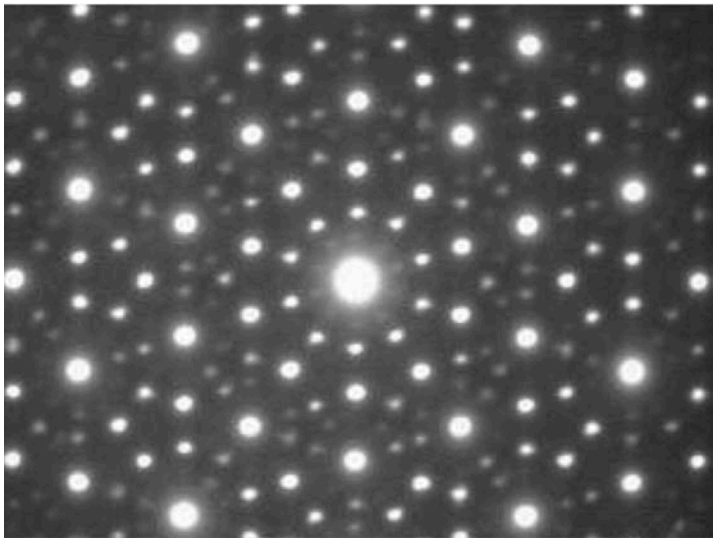
The projection method is due to Paul Steinhardt of the University of Pennsylvania, while the growing algorithmic approach was worked out in detail by the British mathematician John Horton Conway and Roger Penrose himself.

[5] Penrose received the Nobel prize for Physics in 2020, not for his 'tilings' but for 'his discovery that black hole formation is a robust prediction of the general theory of relativity.'

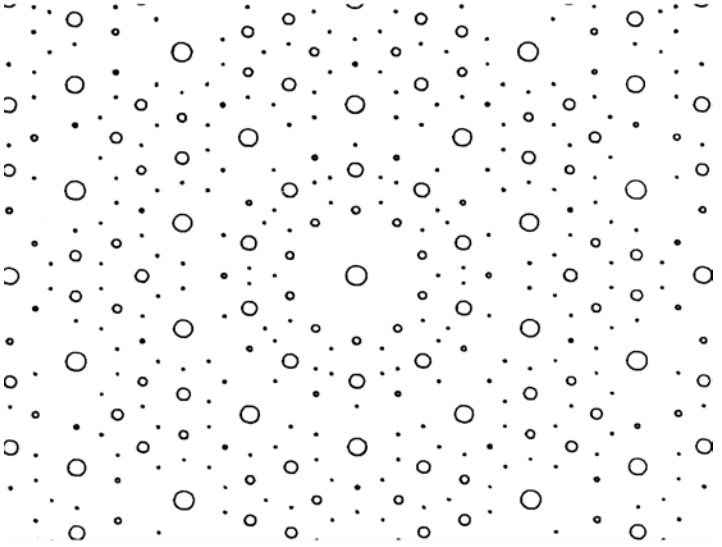
(a) A quasi-periodic tiling of the plane with fivefold symmetry with two types of rhombi. The sharp angles of the rhombi are 72 and 36 degrees.



(b) The (Penrose) quasi-periodic tiling (P3) of the plane with a *'local'* fivefold symmetry. It is possible to completely cover the plane by this arrangement with only two different types of tiles.



(c) The diffraction pattern of a quasicrystal (the $\mathrm{Al}_1 6\mathrm{Mn}$ alloy) having a fivefold symmetry.



(d) The calculated diffraction pattern from a projected higher-dimensional lattice, in a direction orthogonal to a fivefold axis (as in Figure (b)).

Figure III.2.26: A quasicrystal with fivefold symmetry.

Further reading.
On condensed matter physics:

- *Introduction to Solid State Physics*
  Charles Kittel
  Wiley (2004)
- *Solid State Physics*
  Neil W. Ashcroft and N. David Mermin
  Thomson Press (2003)
- *Principles of Condensed Matter Physics* P. M. Chaikin and T. C. Lubensky
  Cambridge University Press (1995)
- *Modern Condensed Matter Physics*
  Steven M. Girvin and Kun Yang
  Cambridge University Press (2019)

On quasicrystals:

- *Quasicrystals: The State of the Art*
  D.P. di Vincenzo, P.J. Steinhardt
  World Scientific (1991)
- *Quasicrystals and Geometry*
  M.Senechal
  Cambridge University Press (1995)

Chapter III.3

The electron collective

Bands and gaps

Electron states in periodic potentials

Two limits. If the nuclei are positioned on the sites of some regular cubic or hexagonal crystal lattice, the electrons no longer move in a spherical electric field of a single nucleus which would give rise to the atomic bound state orbits, rather the electrons experience a periodic electric potential due to the nuclei on the lattice. You may imagine some set of energy wells with a characteristic depth $-V_0$ separated by a distance a. To get an idea of what may happen in this situation we can approach it from two sides as I indicated in Figure III.3.1.

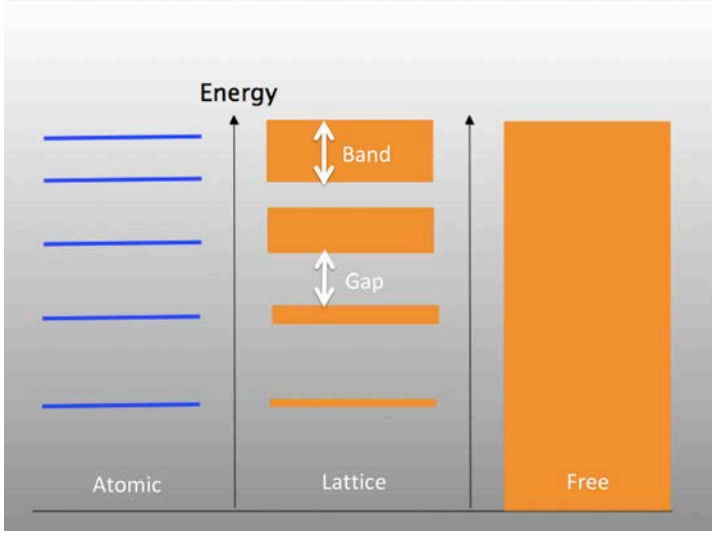


Figure III.3.1: Energy levels, bands and gaps. For individual atoms (l), free electrons (r), and for a periodic lattice of ions (m).

The first approach starts on the left-hand side where we assume that the separation a of the nuclei on the lattice would be large compared to the sizes of the electron clouds of the individual atoms. Then the electron states stay localized around each atom and would maintain the typical atomic spectrum as given on the left. For a solid of N atoms each level would be N-fold degenerate. Now if we start making the separation a smaller, then at a certain point the clouds of ing atoms would start overlapping, and the electrons would start feeling each other's presence due to both their charge and the exclusion principle. This repulsion would deform the clouds and therefore the energy levels would start to split. As a consequence energy bands of narrowly split levels start showing up in the spectrum as indicated on the diagram in the middle.

We could also approach the problem from the right-hand side where we start with V_0 small. Then we would just have the spectrum of free electrons moving through space, and these can have any energy. In other words, the spectrum is continuous as indicated in the diagram on the right. If we let the potential barrier grow, energy gaps would open up and we would again end up with the spectrum in the

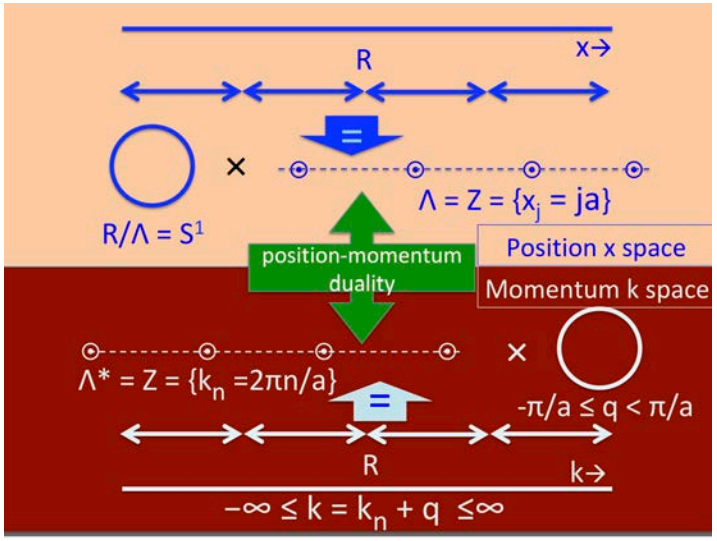


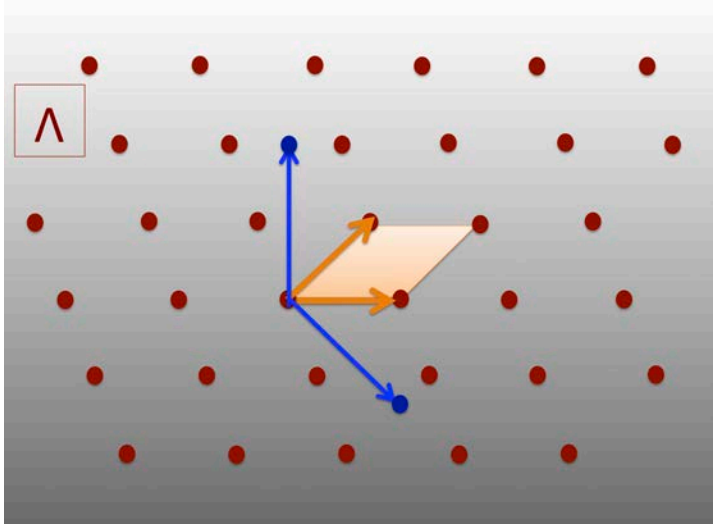
Figure III.3.2: *Position-momentum duality.* In this figure we explain the real space- momentum space duality in the case of a periodicity with period a in the potential for example. We know position space we have $R = -\infty \leq x \leq \infty$ while the free particle momenta are also unbounded $R = -\infty \leq k \leq \infty$. Working top down: (i) divide x-space up in identical pieces of size a which are periodic so we can think of them as little circles (ii) relabel the coordinates a map $x = ja + \varphi a/2\pi$. on a pair (φ, j) where $S^1 = 0 \leq \varphi < 2\pi$ is the angular coordinate of a circle with radius $a/2\pi$, and $j \in Z$ an integer with $-\infty \leq j \leq \infty$.

middle. So coming from the left it is bands that form and coming from the right it is gaps that open up.

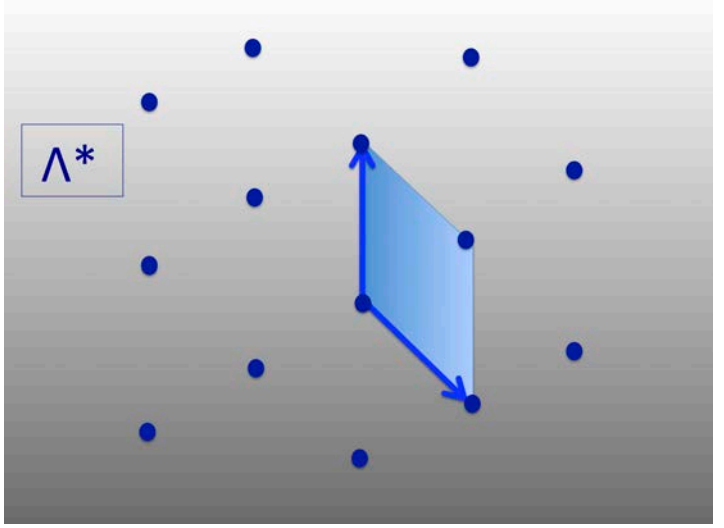
Periodicity and the reciprocal lattice. Let us consider the one-dimensional case where the electrons will move in the periodic potential of the ions on a lattice. The periodicity implies an invariance of the potential under translations over the lattice distance a. And the electron wavefunctions will then carry certain representations of that symmetry. The fact that the potential is periodic does *not* mean that the wavefunctions themselves have to be periodic. The situation is similar to the case of the single atom where the potential is spherically symmetric around the nucleus, but the quantum states are generally *not* spherically symmetric. They form representations of the rotation group labeled by the quantum numbers l and m.

The situation we have depicted of the right hand-side of Figure III.3.1 is illuminating. Let us consider the free particle limit of the spectrum and think of them as states in a periodic (though vanishing) potential. This perspective is visualized in Figure III.3.2 where the top and bottom half are dual to each other. We start in ordinary position x-space which in one dimension is just the real line R. We think of it as a periodic sequence of intervals of size a, the lattice distance. This means that we interpret periodic x-space as a product of a circle with circumference a and a infinite lattice $\Lambda = Z$ with points x_j labeled by an integer j and where $x_j = ja$. So we may now quantize the free particle on this product space, and try to recover the free particle spectrum on the real line, being a continuous spectrum $-\infty \leq k \leq \infty$, as indicated by the real line at the bottom of the figure. The free particle quantization on the circle of radius a yields states that correspond to the discrete 'reciprocal' lattice $\Lambda^* = Z$, labeled by set of integers $\{-\infty \leq n \leq \infty\}$ and corresponding $k-$values $k_n = 2\pi n/a$. It is strictly analogous to the simple Bohr atom. The quantization of a discrete position lattice produces states labeled by a continuous set of values q that form a circle, a periodic interval $-\pi/2 \leq q < \pi/a$. This fundamental domain of $q-$values is called the first *Brillouin zone*. Combining these plane wave quantum numbers we indeed recover the overall k spectrum by simply multiplying the individual exponential (wave functions) which leads to the identification: $k = k_n + q$, corresponding to adding the exponents.

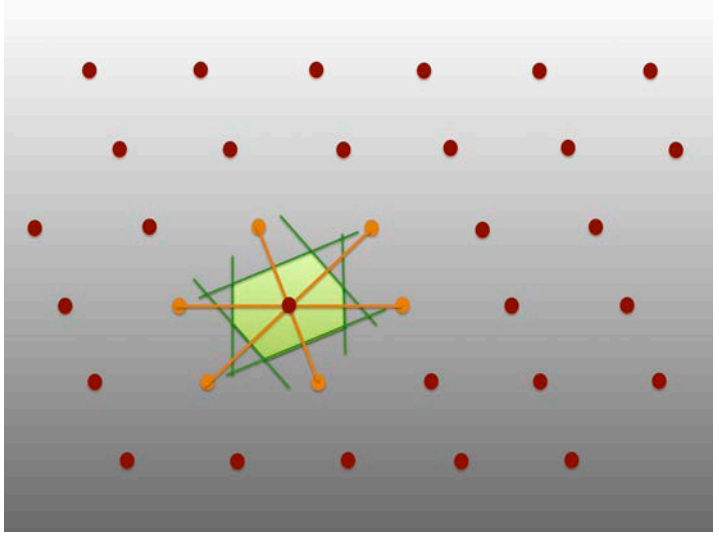
The Brillouin zone. The procedure just outlined is actually quite general, and works in any dimension. You start with an d-dimensional periodic lattice Λ in R^d where we basically identify the points of the x-lattice. This means that the space R can be thought of as a 'product' of a d-dimensional torus R/Λ times the lattice Λ. Free particle quantization gives then a part from the torus which yields



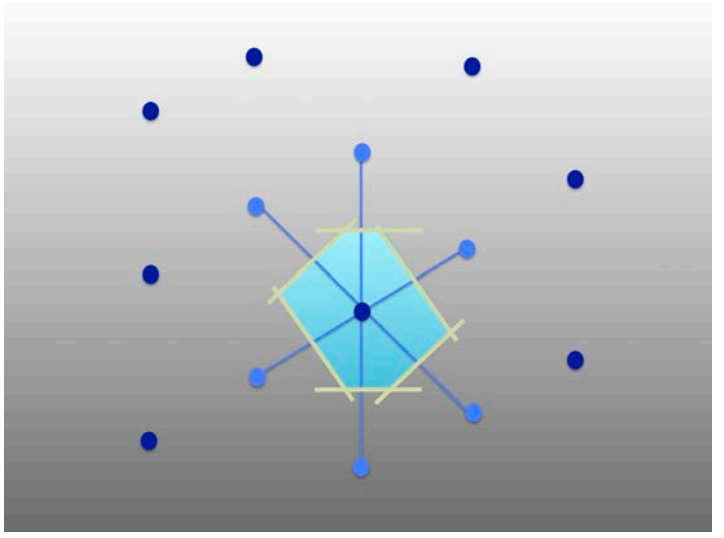
(a) The lattice Λ in x-space.



(b) The dual (reciprocal) latice Λ^* in k-space.



(c) The Wigner-Seitz cell.



(d) The (first) Brillouin zone.

Figure III.3.3: The real space lattice and the reciprocal wave vector lattice.

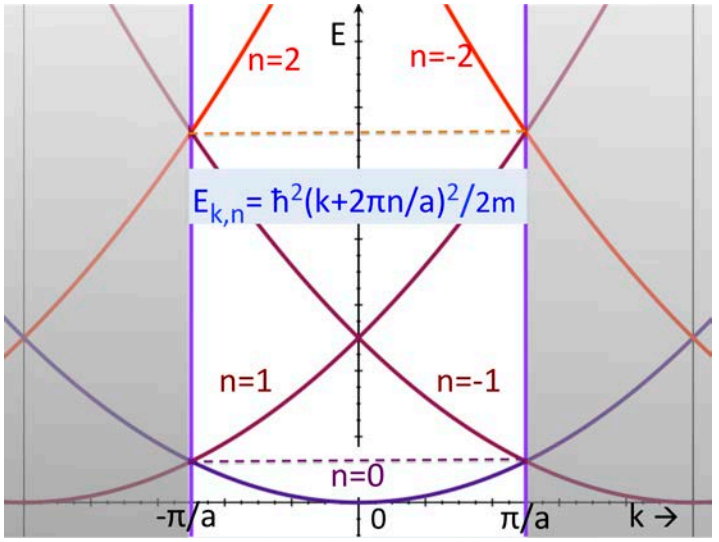


Figure III.3.4: *The Brillouin zone.* We have plotted the energy as function for momentum for free electrons (in one dimension) but have shifted the momentum by an integer times the smallest reciprocal lattice vector $k_1 = 2\pi/a$ as to bring it in the Brillouin zone $-\pi/a < k < \pi/a$, the white colored region. The horizontal axis is the momentum axis, along the vertical axis we have put the electron energy $E = E_n(q)$.

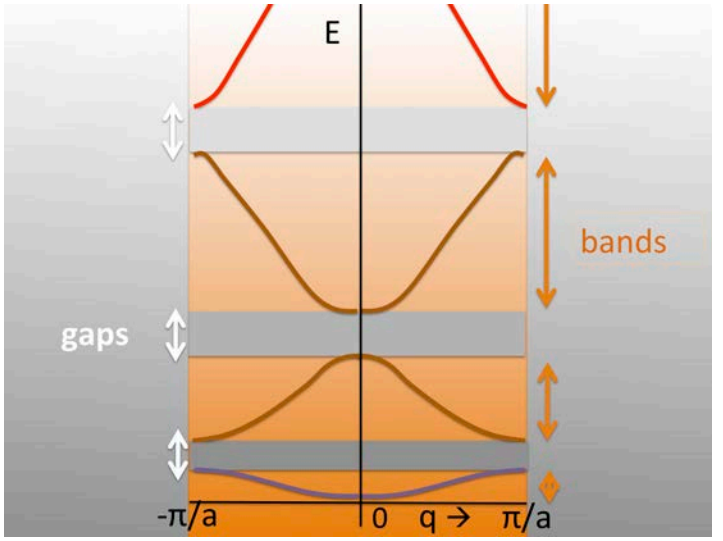


Figure III.3.5: *Gaps open up.* Gaps open up where dispersion curves cross the boundary of the Brillouin zone, or where they intersect. Even though the states will be deformed, the label n of the previous figure remains the label for two successive bands.

the dual or reciprocal lattice Λ^* and a part from the lattice which produces a particular dual torus.

We have illustrated this explicitly for the two-dimensional case in Figure III.3.3. In the Figure (a) we highlighted the so-called periodic *unit cell*, where the symmetry group of Λ is generated by the two basic orange translation vectors. In fact there is an even smaller so-called *fundamental domain* with which the whole plane can be tiled through periodic copying. This domain is highlighted in Figure (c) and obtained as follows. First we start at the origin, and connect it with all ing sites (orange lines), then we draw the perpendicular bisectors (green lines) of the connecting lines. These bisectors then enclose a fundamental periodic (closed) domain called a *Wigner-Seitz cell*. One easily verifies that this cell allows for a space-filling tiling. In figures (a) and (b) we show the construction of the dual lattice, the vectors in the lattices have to satisfy the duality condition:

$$e^{i\mathbf{k}_n\cdot\mathbf{x}_i} = 1\ ,\ \mathbf{k}_n \in \Lambda^*\ ,\mathbf{x}_j \in \Lambda\ . \qquad \text{(III.3.1)}$$

The basic translation vectors defining the reciprocal lattice $\mathbf{T_1}$ and $\mathbf{T_2}$ are obtained from the basic translation vectors $\mathbf{t_1}$ and $\mathbf{t_2}$ by the conditions $\mathbf{t}_i \cdot \mathbf{T}_j = 2\pi\delta_{ij}$. The fundamental domain of the dual lattice constructed in Figure (d) is by condensed matter physicists referred to as the (first) Brillouin zone. The 'Brillouin zone' is the 'Wigner-Seitz cell' in wave-vector space.

Electron wavefunctions: bands and gaps. Let us return to the one-dimensional case, and look at the states in the free particle limit as we have depicted in Figure III.3.4. We have plotted the energy as function of the momentum, or the dispersion $E = E(k)$, but we reduced the k-value by some dual lattice vector $2\pi n/a$, as to bring it in the Brillouin zone. In other words we plot $E(k) = E_n(q)$, and that is in fact what is shown on the right-hand side of Figure

III.3.1, and in the parametrization given in the lower half of Figure III.3.2. So indeed, the free electrons can have any energy $E_n(q) \leq 0$. Note that in the resulting spectrum the levels fold over at the boundaries $(q = \pm\pi/a)$ and cross in the middle where $q = 0$. If we return to Figure III.3.1 we have argued why by increasing the nuclear potential the continuous spectum will break up, and gaps will open up as depicted in Figure III.3.5, exactly for the special values of q as indicated. Let us now after this introduction move on to the generic spectrum of the quantum electron fluid in an ordinary solid.

Valence and conduction bands. In Figure III.3.6 we give the band structure in the periodic potential landscape of the lattice in which the electrons live. The landscape is characterized by the interatomic distance and the height V_0 of the potential barrier. The electrons fill the bands to a certain maximum level which is called the *Fermi level*, marked by the white dashed line. The two bands closest to the Fermi level are called the *valence band* and the *conduction band* and as we will see the properties of the material will depend strongly on where these bands are located with respect to the Fermi level. The inner electron bands below the valence band consist of pretty much localized states. The allowed states in the 'conduction' bands are not localized but extended, which means that electrons move anywhere in the sample.

Conductors and insulators. How the electrons in the solid collectively behave strongly depends on the position of the *Fermi surface*, which Figure III.3.7 demonstrates. If the Fermi level is in the middle of the valence band, the electrons can move easily because there will be many states available with some more energy, and the material is therefore a *conductor* for electric currents. If the valence band is completely filled and there is not energy enough to enter the conduction band, the electrons cannot move, and we are dealing with an *insulator*. We say that the medium has an *energy gap* – is *gapped*. The intermediate case of a *semiconductor* deserves a section of its own.

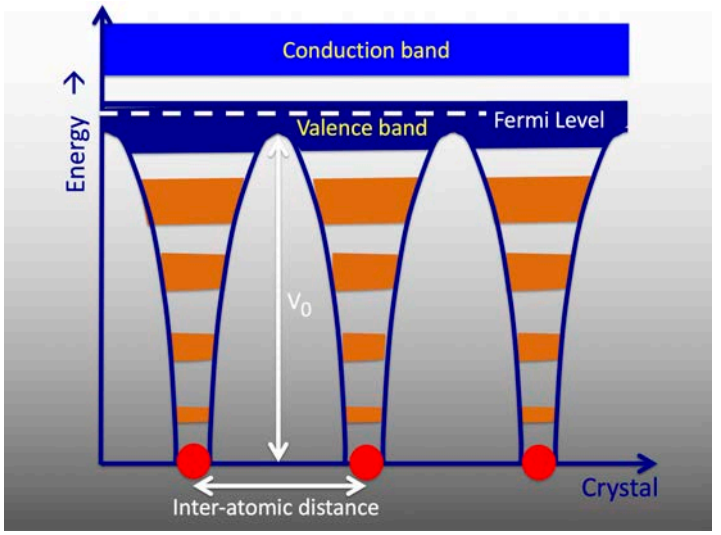


Figure III.3.6: *Electron bands in a crystal.* In a crystal the energy levels are all filled up to the Fermi level (dashed line). The two bands closest to the Fermi level are called the valence and conduction band. The periodic potential is characterized by the interatomic distance and the height of the potential barrier V_0.

Semiconductors.

Finally we can imagine that the energy gap between valence and conduction band is narrow, so that not much energy is needed to excite electrons into the next band. This is typically the situation in a *semiconductor*. The image on the right in Figure III.3.7 shows a narrow band gap of a semiconductor at room temperature. The coloring indicates that because of the thermal energy some electronic states at the bottom of the conduction band will be occupied leaving some holes in the valence band. In the next figure we show again the typical energy landscape of what is called an *intrinsic semiconductor*, with the two bands and the Fermi level right in between. The electron/hole density in equilibrium is determined by the energy difference between the (conduction/valence) band edge and the Fermi level, which means that as $E_- = E_+ = E_G/2$ the number of charge carriers $n_\pm$ is exponentially suppressed

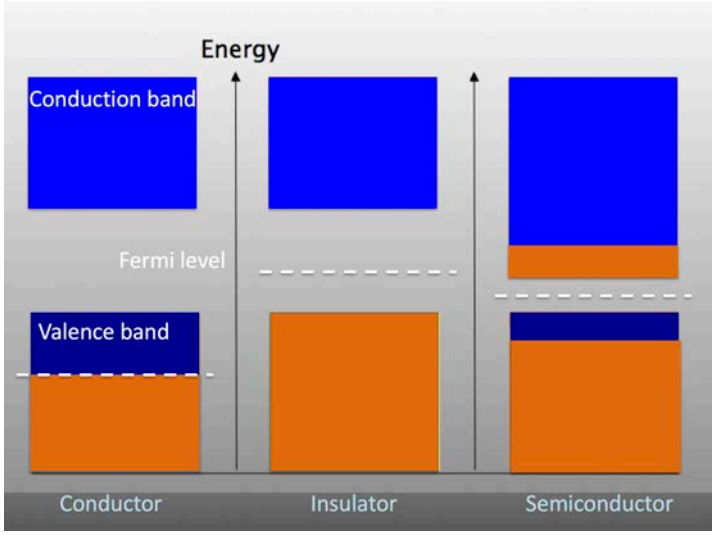


Figure III.3.7: *Energy bands.* The admissible energy levels for the electrons form the *valence* and *conduction* band, where the Fermi level is marked by the dashed white lines. We distinguish a *conductor* (l) where there is basically no gap, an *insulator* (m) where the valence band is filled and there is a big gap, and a *semiconductor* (r) with a narrow gap. The filled states are colored orange and empty states blue.

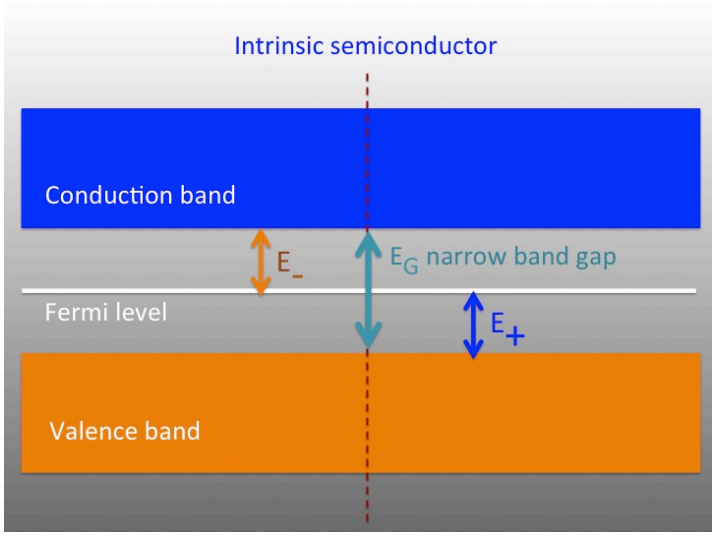


Figure III.3.8: *The intrinsic semiconductor.* The intrinsic semiconductor is characterized by a narrow gap between valence and conduction band, with the Fermi level exactly in between. The horizontal axis is the space axis, along the vertical axis we have put the electron energy.

by a Boltzmann factor $\exp(-\mathrm{E_G}/2\mathrm{kT})$. But this also implies that its dependence on the energy gap is exponential and that fact is exploited in the idea of doped semiconductors on which all basic semiconductor devices such as transistors are based.

Semiconductors like silicon are at the heart of all modern information storing and processing devices. It is not by accident that the Californian cradle of the information revolution we have witnessed is called 'Silicon Valley'. And it was because of the ever smaller scales at which the semiconductor switches (transistors) could be implemented and exploited that the spectacular large-scale integration of processor and memory chips became possible.

A doped semiconductor. The possibility of *doping*, allows you to somewhat customize the energy landscape in semiconductors. What one does is to replace a certain percentage of the silicon atoms in the lattice by either phosphorus (P) or boron (B) as indicated in Figure III.3.9. In the periodic table phosphorus is the right-hand neighbor of silicon and therefore provides an extra electron, which makes the material somewhat more negatively charged. The effect is to basically lower the band energies with respect to the Fermi level. Substituting with boron has the opposite effect, as boron sits in the column to the left of silicon, and therefore has one valence electron less; the semiconductor will have an excess of positive charges or holes. One may also dope the opposite sides of a semiconductor differently, in which case one gets a *pn-diode* or *pn-junction*, as we have depicted in Figure III.3.10. In addition to the band gap $\mathrm{E_G}$, a new energy scale $\mathrm{E_D}$ is introduced by the doping: on the left side we have many electrons and on the right side only a few, because there is a relative suppression factor $\exp\left(-\mathrm{E_D}/\mathrm{kT}\right)$. For the holes the story is just the opposite, many holes on the right and

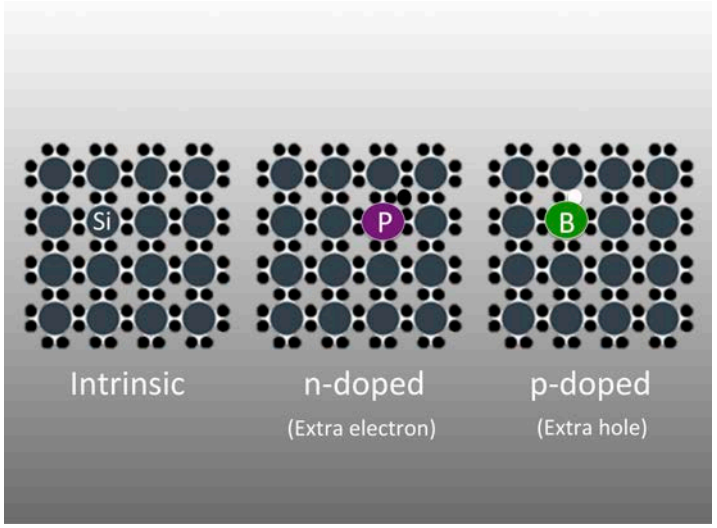


Figure III.3.9: *Doped semiconductor.* We can replace a certain fraction of the silicon atoms in the lattice by either phosphorus (P) or boron (B). The former yields an excess of negative charge carriers (electrons), called *n-doping*, whereas the latter leads to an excess of positive charge carriers (holes), called *p-doping*.

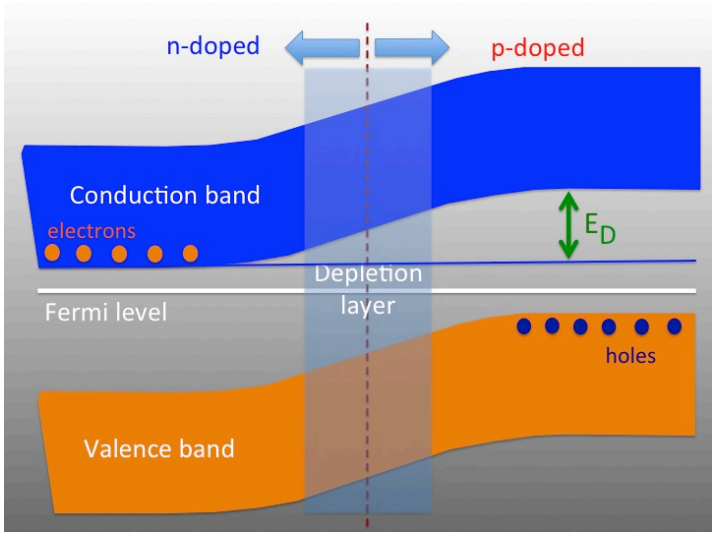


Figure III.3.10: *pn-junction.* By doping a semiconductor we can shift the band structure. With an excess of negative charge carriers (n-doping) we lower the bands, whereas with an excess of positive charge carriers (p-doping) the the bands move up in energy. In the figure you see the band profile of a *np-doped semiconductor* or a *pn-junction*.

few on the left. In the middle in the so-called *depletion layer* there are neither free charges nor free holes, it acts as an insulating layer. The Fermi level is the same on both sides, as you can always briefly shortcut the external wires till this equilibrium is established.

Two semiconductor devices. This pn-diode is a simple and useful semiconductor device. Let us briefly indicate two applications without going into much detail.

The photo-voltaic cell. The first possible application is to make a photo-voltaic cell which basically turns solar radiation in the form of photons into electron hole pairs by just exciting electrons from the valence band to the conduction band. This is illustrated in Figure III.3.11, and amounts to creating an opposite charge excess on both sides of the device. In other words creating a voltage difference between the two external plates. Clearly if we couple enough of them in a big array, we can generate high voltages and big currents. And this is a common way to convert solar radiation into electric power. The challenge is to make the efficiency large enough, so light has to be able to enter the semiconductor sufficiently as to maximize the absorption.

The Light Emitting Diode (LED). In Figure III.3.12 we show what happens if we connect the leads to a battery where we introduce a third independent energy scale $\mathrm{E_B} = eV_\mathrm{B}$. The battery induces an energy (voltage) difference corresponding to $\mathrm{E_B}$ between the left and right Fermi levels. These levels split near the depletion layer. One can imagine what happens, the negative lead pushes the electrons from the left towards the junction, and similarly the positive lead will push more holes in the system from the right. The effect is that the depletion layer becomes narrower and in fact if the voltage is high enough you will get a current of electrons and holes through the junction. However, as in

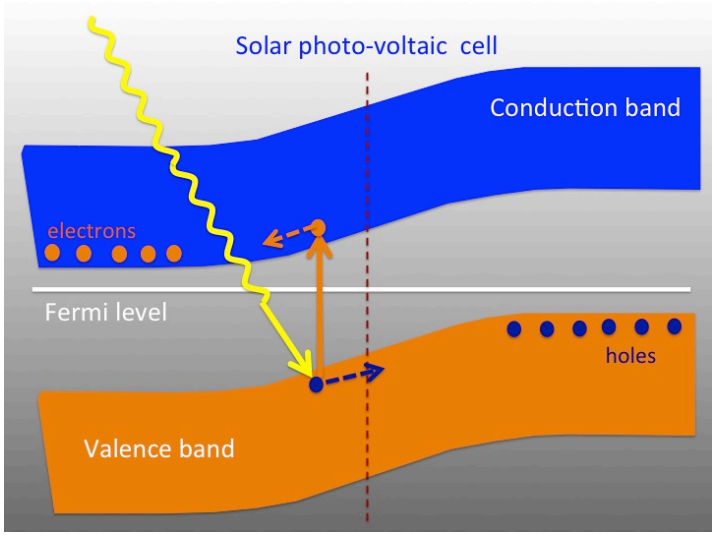


Figure III.3.11: *Photo-voltaic (Solar) cell.* If we have a transparent np-doped semiconductor, light (photons) can be absorbed by the electrons in the valence band and be excited to the conduction band leaving a hole behind. So a voltage will build up over the cell and a current can flow.

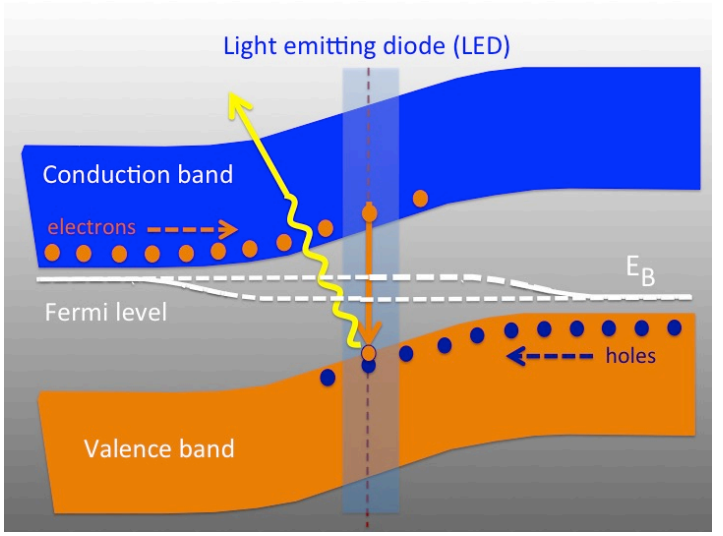


Figure III.3.12: *Light emitting diode (LED).* The LED is more or less the converse of the photo-voltaic cell, in that we now apply a voltage over the semiconductor, which changes the Fermi level on the negative/positive sides. This leads to a recombination of electrons and holes in the center region of the junction producing light.

a stationary state, the relative charge densities between the right and left have to remain exponentially different. What happens is that in the middle region the electrons and holes will recombine and that produces radiation that may be absorbed in the material, but of course it is also possible to implement this in a way that the radiation in the form of photons escapes, and we have a LED. It is a clear advantage that the energy is directly converted into electromagnetic energy, not by heating a wire which in turn starts radiating. Voltages and currents can therefore remain quite low as long as a sufficient percentage of recombined pairs results in visible photons. At present the differences are quite stunning: the LED has a lifespan that is about a factor 50 higher than that of an incandescent bulb, while it costs about a factor 30 more. It is the energy consumption that makes the big difference, because that provides an additional factor of 60. This means that over the lifetime of an LED your yearly electricity bill would be reduced by a few hundred euros/dollars! These numbers also underscore the relative waste in the form of heat that is produced by the old-fashioned light bulb.

Superconductivity

Phonons. It is exciting to go one step deeper into possible scenarios for the collective behavior of the electrons. Looking more closely at the lattice, we know that the nuclei cannot be completely fixed at their positions on the lattice. They are subject to quantum and thermal fluctuations and these lattice fluctuations lead to waves propagating through the lattice, which are just the familiar sound waves as a matter of fact. In the quantum perspective these waves are considered to be *quasi-particles* which are called *phonons*. So where photons are complementary to light waves, so are these phonons complemen-

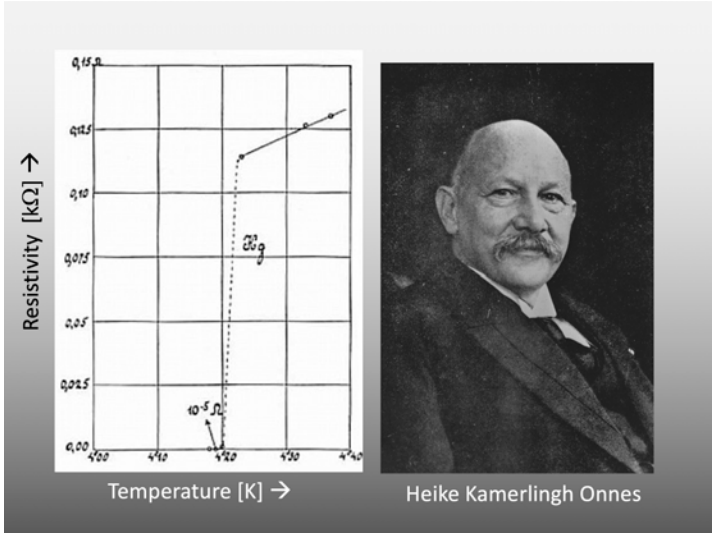


Figure III.3.13: *Superconductivity.* The discovery of superconductivity, as the measurement of a sudden dramatic drop in resistivity of solid mercury, was made in 1911 in Leiden by Heike Kamerlingh Onnes. It took more then fifty years before a fundamental understanding of this phenomenon was achieved.

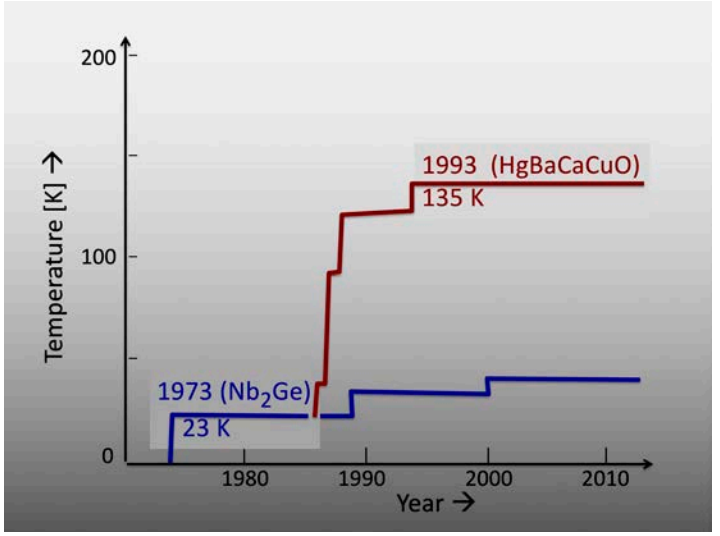


Figure III.3.14: *High temperature superconductivity.* The maximum temperature at which superconductivity takes place has increased dramatically during the last quarter of the 20th century, but appears to have stabilized again. A fundamental understanding of the underlying mechanism, however, is still lacking.

tary to sound waves, and because sound only propagates through a material medium these quasi-particles are not really fundamental, they are quantized collective excitations of the underlying medium.

Cooper pairs. Now the oscillating nuclei are charged and we should expect that these waves interact again with the electrons. In particle language the phonons will couple to the electrons. And the interesting feature of these interactions is that they lead to an effective attractive force between the electrons. In other words, the 'phonons' become the carriers of an attractive force between the electrons. What happens is interesting, close by the electrons are repelled because of their charge, but that repulsion is screened on larger distances and there the attractive force due to the phonons becomes dominant and creates bound states of electrons, the electrons pair up and form so-called *Cooper pairs*. At low temperatures you may think of the Fermi surface as a sphere in momentum or k-space with well defined radius $\mathrm{k_F}$. A Cooper pair is formed by two electrons at opposite points of the sphere, where furthermore the electrons have spins pointing in opposite directions. In Figure III.3.15 we have indicated the Fermi sphere with two Cooper pairs at the surface, each pair bound through the exchange of a virtual phonon. So we should think of the electron collective no longer as a community of singles but of couples and once more that strongly affects the states that are allowed just as in our earlier societal analogue.

The superconducting ground state. I have already referred to the spin of particles and the Pauli exclusion principle, which decrees that two half-integral spin particles cannot occupy the same state whereas integral spin particles can. But after the electrons pair up, we are no longer dealing with a collective of spin 1/2 electrons, but with pairs of electrons with opposite spins, which means that the pairs have spin zero. And that has dramatic conse-

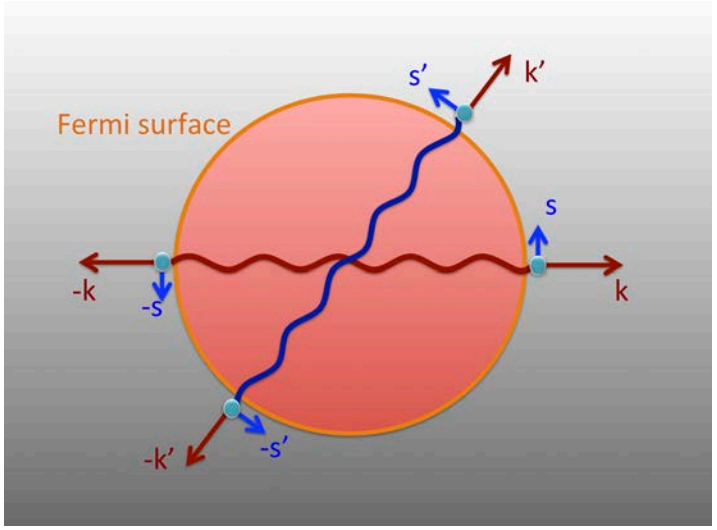


Figure III.3.15: *Cooper pairs.* Cooper pairs are bound states of two widely separated electrons caused by the exchange of virtual phonons. This turns the electron collective effectively into a gas of charged bosons which then condense into the superconducting BCS state.

quences: whereas the electrons cannot sit in the same state and push each other to ever higher and higher energy states, the charged bosonic pairs all can sit in the same lowest energy state. You can imagine that there is an enormous energetic advantage for the system of electrons to pair up and all 'condense' in the ground state. Well, it does happen, and we see that for certain conductors, if we cool them down sufficiently, the pairs can form and condense into a surprising new state of matter: the material becomes superconducting. A *superconductor* is a conductor with the miraculous property that it conducts electricity with absolutely *zero* resistance! The most dramatic fact is maybe that this phenomenon is a macroscopic manifestation of quantum theory, *the superconducting state is a macroscopic quantum state*. This is possible because all the Cooper pairs have condensed into a single quantum state.

Bose-Einstein condensates. These kind of condensation effects are a manifestation of Bose-Einstein condensation an effect predicted as early as 1924 by the Indian physicist Satyendra Nath Bose and Albert Einstein. And indeed, many other examples have since been found: for example He^4 is a boson and therefore can condense at very low temperature in a state that exhibits the amazing property of *superfluidity*. As we discussed in the previous chapter, there is no viscosity in a superfluid: another one of these quantum miracles which would be inconceivable from a classical point of view. The Bose-Einstein condensates which have been observed in diluted atomic gases, and for which the Americans Eric Cornell, Carl Wieman and Wolfgang Ketterle received the Physics Nobel prize in 2001, are another recent discovery. These condensates are close to the theoretical setting described in the original papers of Bose and Einstein.

Some history. We have made a small *tour d'horizon* to give you a sense of how rich and surprising the macroscopic behavior of a collective of atoms may be, and how intricate the balances of forces are, and to what kind of exotic properties of materials this may lead. It also shows how creative one has to be to get to a detailed physical understanding such exotic properties. It is worth pointing out that superconductivity was discovered by Heike Kamerlingh Onnes in Leiden as early as 1911. He found that the resistance of solid mercury immersed in liquid helium suddenly dropped to zero at a temperature of 4.2 K, as shown in Figure III.3.13. The story goes that he generated a persistent circular current and managed to take it along to Amsterdam to show it to his colleagues over there! Kamerlingh Onnes received the Nobel prize in Physics in 1913 for 'his investigations on the properties of matter at low temperatures which led, *inter alia*, to the production of liquid helium.'

The microscopic mechanism underlying superconductivity remained a complete mystery for a long time. The Russian physicists Lev Landau and Vitaly Ginzburg proposed an effective field theory explaining quite a lot of the phe-

Figure III.3.16: *Magnetic levitation.* A little magnet will be lifted above a superconductor, because of the Meissner effect, which means that magnetic field lines are expelled from a superconducting region. The aura of magic is caused by the boiling liquid nitrogen needed to cool the high-temperature superconductor. (Source: Michigan State University.)

nomenology of the superconductors, but it was not until 1957 that the fundamental quantum mechanism including the pair formation and the precise structure of the superconducting ground state was put forward by the American physicists John Bardeen, Leon Cooper and Robert Schrieffer, who received the Nobel prize for their groundbreaking work in 1972. This splendid theory is known as the BCS theory of superconductivity.

Ever since the 'BCS' breakthrough in the understanding of superconductivity there has been a host of detailed quantum mechanical explanations for the highly surprising ways collectives of atoms may behave and turn into molecular gases, liquids, glasses, liquid crystals, magnets, superconductors or Bose-Einstein condensates, or even assemble into large molecules, all depending on the parameters of the theory. Again, this is the branch of physics which Philip W. Anderson, the celebrated American condensed matter theorist who died in 2020, characterized by the credo 'more is different', referring to the splendid diversity of collective quantum behavior that emerges in macroscopic systems consisting of many interacting constituents. We have emphasized that the differences cannot always be traced back to the differences in the constituent particle types. Though the type of interactions these have is absolutely crucial, the macroscopic phase that is realized may also depend on external parameters, like the temperature, the density, the presence of a magnetic field and so on. To conclude we may say that in trying to understand and predict the splendid diversity of emerging properties, quantum reasoning has become absolutely indispensable.

The Meissner effect. You might wonder what happens if we apply a magnetic field to a superconductor. This is an interesting question to ask because we know that a conductor tends to counteract a change in the magnetic field, which means that currents are generated which are such that they generate a field in the opposite direction. Now you can imagine that because there is no resistance in the superconductor these currents will keep running thereby permanently counteracting the change in magnetic field. The net result is remarkable: magnetic fields cannot penetrate a superconductor! This expulsion of magnetic fields from superconducting regions is called the Meissner effect, after the German physicist Walther Meissner who discovered it in 1933.

Here some qualifications must made though. The first is that if we keep increasing the magnetic field we end up breaking the pairs and the superconducting phase is destroyed. The second is more interesting and follows because the electrons (and pairs) have a funny property. It turns out that they cannot detect a specific amount of magnetic flux. What happens in the so-called Type II superconductors is that the magnetic flux can enter the superconductor if it is in quantized portions the electron pairs can't see. In other words, there is a minimal unit of magnetic

flux Φ_0 that is compatible with a condensed charge q and it is given by the simple relation $\Phi_0 = 2\pi\hbar/q$.

In a three-dimensional superconductor these magnetic flux lines that enter the superconductor line up parallel to the direction of the external magnetic field. However, the flux lines repel each other and therefore if you increase the strength of the field and look in a plane perpendicular to the field you see that they tend to form a nice triangular lattice. I should point out an additional or better complementary view on this situation. The fact is that in the core of these magnetic filaments the medium becomes a normal conductor again. So, in a sense you can say that the magnetic field did not enter the superconductor after all but corresponds to filaments of a normal conductor in the superconductor.

I have all along been emphasizing the use of symmetry arguments. What about the superconducting phase, are they of any use there? The answer is affirmative. Though the argument is somewhat more complex. We all know that electric charge is conserved: you cannot lose an electric charge; it may be transferred from one fundamental particle to another, for example in reactions like proton + electron goes to neutron etc. We have mentioned in previous chapters that this conservation law is a consequence of the internal symmetry called *gauge invariance*.

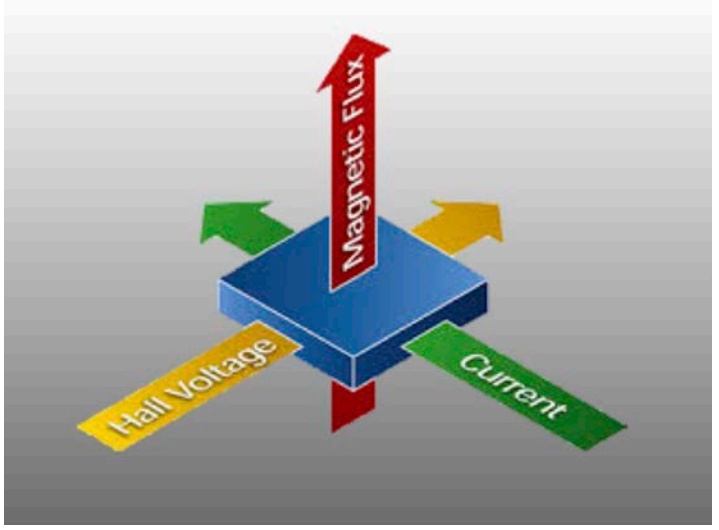
But if, like in the superconductor, the groundstate is filled with electrically charged particles, then the electric charge is no longer conserved, you can change it by arbitrary multiples of $2e$ without changing the physical situation. The point is that the superconducting state is unusual in that there is no definite number of electrons or pairs in that state. So, the story here is that in the superconducting phase charge is no longer conserved because the gauge symmetry is broken. But if a symmetry is broken then we must ask whether there are not defects that we have to take into account. Yes indeed, the defects are precisely the magnetic vortex lines we have been discussing. The symmetry breaking story once more fits exactly the phenomena observed.

The quantum Hall effect

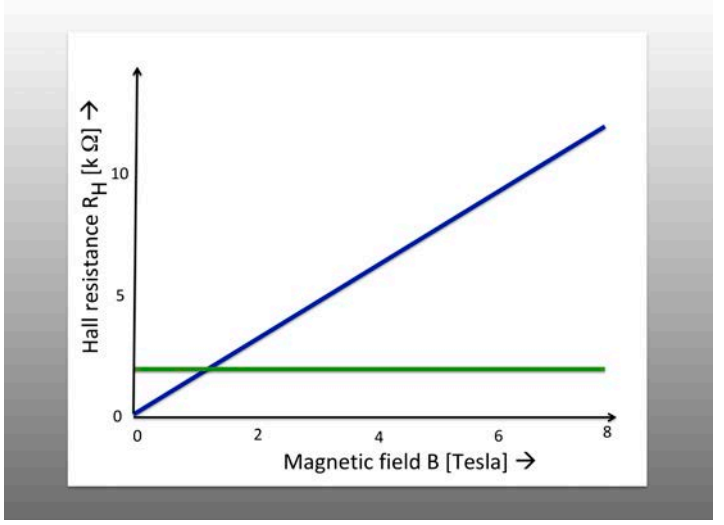
In the phenomenon of superconductivity we have seen one of the more subtle ways the system of a rigid lattice can interact with the gas of electrons and give rise to a rather surprising form of collective behavior. Are there other examples of interactions electrons may engage in that drastically change their collective behavior? I wouldn't ask you if the answer wasn't yes. A stunning example is the so-called *quantum Hall effect*: it occurs just like superconductivity and superfluidity only at temperatures of a few Kelvins so that its applications have been limited so far. The setting for the quantum Hall effect is a two-dimensional conductor (imagine for example a conducting boundary layer between two insulators) where we apply a strong magnetic field perpendicular to the surface. This situation is depicted in Figure III.3.17(a).

The physics in this setting is rather counterintuitive. Imagine a little slab of quantum Hall medium and applying a voltage difference V in the x direction. In a normal conductor a current I would start flowing in the x-direction according to Ohm's law decreeing that $I = V/R$ so, inversely proportional to the resistance R. In the quantum Hall medium however, the current starts flowing in the y direction, perpendicular to the applied field! This is even the case in classical physics as Edwin Hall already discovered in 1879. The transversal Hall resistance as a function of the applied magnetic field (with fixed current) is plotted in Figure III.3.17(b). We talk about a transversal or Hall-resistance (ρ), and a Hall-conductivity $\sigma = 1/\rho$.

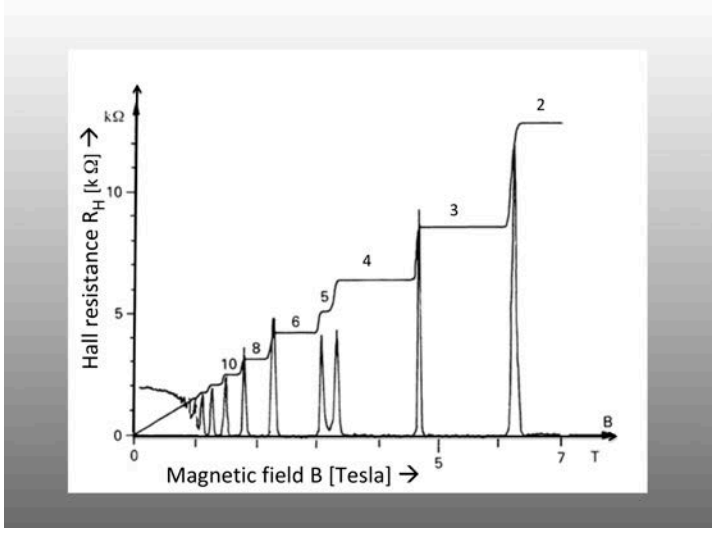
The integer quantum Hall effect. To consider this system quantum mechanically, there are two things that we ought to understand. The first question is the behavior of a single



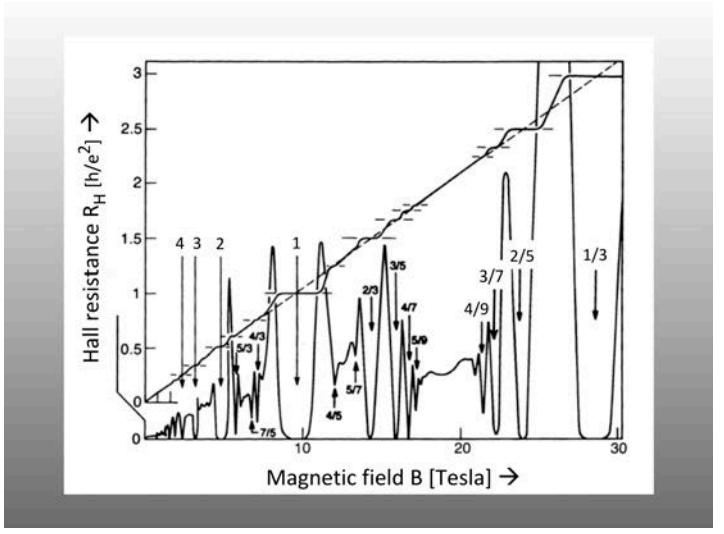
(a) The quantum Hall setup. Driving a current through a planar conductor with a strong magnetic field B orthogonal to the plane yields a transversal potential $\mathrm{V_H}$.



(b) The classical Hall effect shows a linearly rising $\mathrm{V_H}$ (blue line) as a function of the applied magnetic field, while keeping the current constant (green line).



(c) The integer quantum Hall effect showing the plateaus with integer ν values.



(d) The fractional quantum Hall effect with from right to left the plateau values for $\nu = \frac{1}{3}, \frac{2}{5}, \frac{3}{7}, \frac{4}{9}$....

Figure III.3.17: *From the classical to the quantum Hall effect.*

electron in a magnetic field, and the second is the collective behavior of electrons in this setting. It is beyond the scope of this book to drag you through the beautiful reasoning, but even if we had done so, the phenomenon remains quite puzzling and counterintuitive. Where according to classical physics the transverse conductivity must grow *linear* with the applied field, in reality it does not! If you increase the magnetic field the conductivity remains constant over certain intervals and the value of that conductivity is strictly quantized according to the surprisingly simple relation $\sigma = \nu n e^2/2\pi\hbar$, where ν is the *filling fraction* which is defined as the electron density n_e divided by the magnetic flux density n_B in fundamental flux units ($\Phi_0 2\pi\hbar/e$): in other words $n_B = eB/2\pi\hbar$. We have plotted plateaux in the Hall resistance for the integer effect in Figure III.3.17(c). What you see is that as a function of the applied magnetic field it has plateaus where it stays constant until it jumps to the next plateau (with lower n).

The fractional quantum Hall effect, When you turn up the magnetic field to large values like 30 Tesla, plateaus also show up for fractional values of ν like $\frac{1}{3}, \frac{2}{5}, \frac{3}{7}, \frac{4}{9}, \ldots$, in which case we speak of the *fractional quantum Hall effect* as depicted in Figure III.3.17(d). In the fractional quantum Hall effect we have the unusual situation that the charge carriers in the medium are no longer electrons. Rather they correspond to localized collective excitations of the system which carry *fractional* electric charges, such as $e/3$ or $e/5$ depending on which plateau you are.

So, to put it in more pictorial terms: if I would add an electron to a quantum Hall system it would 'fall apart' in a set of fractional charges as displayed in Figure III.3.18. However, you should not think of these charge carriers as some kind of special 'quark-like' particles that make up an electron. No, these fractional charges are carried by well-localized collective excitations, special modes of the electron field in the presence of the magnetic flux. So, these collective excitations are not only charged but they also carry a magnetic flux quantum along with them. The flux quanta are in

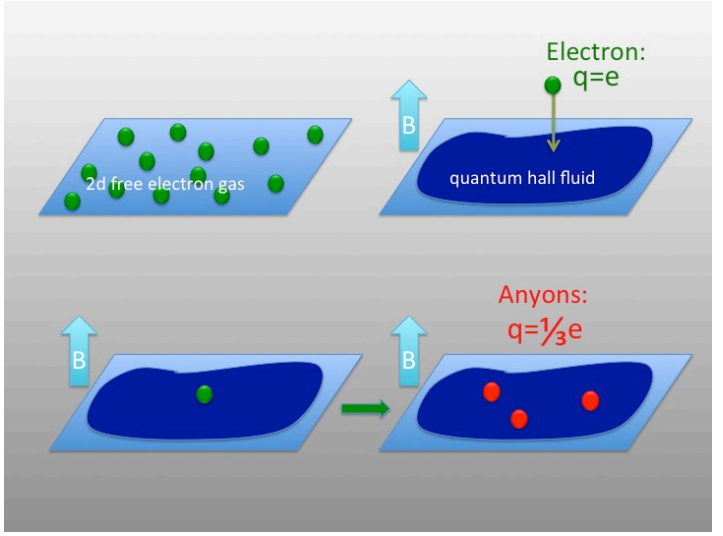


Figure III.3.18: *The quantum Hall fluid.* Putting a 2-dimensional free electron gas near absolute zero in a strong magnetic field one obtains a quantum Hall fluid. Adding a single electron charge to the quantum Hall fluid, the charge will fractionalize into three *anyons* each with charge $e/3$. These anyons are quasiparticles, and are in fact flux-charge composites carrying a exotic spin value $s = q\Phi_0/2\pi = e/3 \cdot \hbar/e = \hbar/3$.

that sense the magnetic defects we saw in the type II superconductors and which become particle like in a plane orthogonal to the magnetic flux, but now these flux particles are dressed with a fractional electric charge. Such dually charged excitations that basically can only occur in two dimensions are called *anyons*. We have discussed such flux-charge composites in the section on spin and statistics on page 405 of Chapter II.5, and more specifically the subsection on two-dimensional exotics on page 416. There we showed that such composites may indeed exhibit not just fractional charge, but also fractional spin and statistics properties. For the case where the basic anyonic charge corresponds to $q = e/3$, we demonstrated that the spin of the anyon corresponds to $s = q\Phi_0/2\pi = \hbar/3$.

Quantum Hall systems are of fundamental interest because they represent truly novel states of matter, the existence of

which nobody had anticipated. The integer quantum Hall effect was discovered by the German physicist Klaus von Klitzing in 1980, for which he received the Nobel prize in 1985, The more complicated fractional quantum Hall effect, featuring the fractional charge and exotic statistics properties was discovered in the early 1980s and a Nobel prize for theory and experiment was awarded in 1998 to Robert Laughlin, Horst Störmer and Daniel Tsui.

Topological order

Quantum Hall conductors constitute entirely novel states of matter, fundamentally different from the more familiar conducting phases, like ordinary conductors, semi- or superconductors, which are usually referred to as Fermi liquids. From 1980 onwards many phases which exhibit similar unusual behavior have been discovered; these phases which are characterized by certain non-trivial topological interactions are now considered manifestations of a generic property called *topological order*. It concerns phases which are *gapped*, which means that there are no massless degrees of freedom in the system, the relevant degrees of freedom are massive like the anyons, and these have topological long range interactions leading to their non-trivial spin and statistics properties.

Quantum statistics. The term *anyon* was coined by the American physicist Frank Wilczek because these fractionally charged particle also have an exotic type of quantum statistics properties. We have emphasized the essential difference between bosons and fermions, where the latter obey the Pauli exclusion principle saying that no two fermions can sit in exactly the same state whereas bosons can. Another way of saying this is that if we consider a multi-particle state and we interchange two identical type particles then the phase of the state may change. In three or more dimensions, if we repeat the interchange operation, denoted by τ, we are back to the original state, so that implies that $\tau^2 = 1$, which means that the phase change has to equal $\tau \simeq \pm 1$. If we interchange two bosons the state remains unchanged $\tau = 1$ and if we interchange two fermions the state changes sign so $\tau = -1$. We have already pointed out that this difference in statistics (we call it statistics because the rule affects the way the particles can be distributed over the available states) accounts for the crucial differences in properties in many body systems. We recall the essential role of the Pauli exclusion principle in understanding the spectrum of atoms with more than one or two electrons.

Braid statistics. The anyons that occur in two-dimensional topologically ordered media satisfy a type of statistics referred to as *braid statistics*, where there is an essential phase difference between interchanging particles clockwise or counterclockwise. So to calculate the state after some time you have to keep track of how often and in what direction the particles have moved around each other. One has to deal with the *braid* of particle world lines in spacetime. And to know the state exactly you have to know the braid. A braid is much like a knot, and the theory of knots is a well-studied subject in the topology of three-dimensional manifolds. If we have a particular braid of five differently colored strands, we could connect the corresponding incoming and outgoing strands to obtain a closed knot made of five strands. It is topological because it doesn't matter at what distance the world lines wind around each other and moreover we may move the strands around and deform the knot; but as long as we don't cut the strands the knot remains topologically the same. The knot will be characterized by a number of topological invariants. In terms of the quantum Hall effect this means the way the quantum state changes only depends on who danced around who and in what order. Another way to say this is that the multi-anyon states exhibit long range entanglement.

All possible braids can be composed of elementary moves of moving neighboring pairs around each other. The set of all such intertwining operations forms again a group,

and the mathematics of such groups is well understood. In higher dimensions one only can have bosons or fermions, exactly because winding the paths clockwise or anti clockwise is topologically equivalent, while in two dimensions there is in principle an unlimited number of topologically inequivalent windings possible and therefore also for the quantum statistics of states. It is even possible that different particle types exhibit non-trivial mutual statistics properties. This means that the phase of the states may change after moving a particle around another type of particle: in other words applying τ^2 to a pair of particles belonging to different species. In general, the multi-anyon states are formally classified as unitary representations of the braid group.

Topological field theory. You may wonder what the theoretical models look like that effectively describe these topologically ordered phases like the quantum Hall fluid. A large and important class of models, but not the only type, are so-called *topological field theories.* In particular, the (2+1)-dimensional *Chern-Simons theories*. It is an effective theory, which describes the phenomenology of the topologically ordered phases to a certain extent. And one must realize that a derivation of this theory from first principles is hard. To give you a flavour of what such theories look like, I show a basic example that is provided by just a (charge q) current $j_\mu =$ coupled to a gauge field A_μ that is described by a $U(1)$ Chern-Simons theory. The equations in relativistic notation are actually quite simple and given by:

$$\frac{\lambda}{2}F_{\mu\nu} = \varepsilon_{\mu\nu\sigma}j^\sigma \;\Rightarrow\; \begin{cases} \frac{\lambda}{2}F_{12} = j^0 \to B = \frac{2\rho}{\lambda} \\ \frac{\lambda}{2}\mathbf{E} = \mathbf{j}_\perp \end{cases}, \qquad \text{(III.3.2)}$$

where the parameter λ is the coefficient of the Chern Simons term, which dependent on the setting will be quantized as well. In the quantum Hall effect it is directly linked to the quantized plateaux conductivity. What these equations imply becomes clear if we look at simple situations:

(i) If there is no charge or current, the equations say that there is no field: this is an expression of the fact that there is a gap, and there are no gauge field quanta maybe because they are too heavy to be excited. In other words, the pure Chern Simons theory has a 'gauge field' but that field does not describe local field degrees of freedom like photons. It is a purely topological theory, meaning that the only physical *observables* are the path dependent phase factors corresponding to closed loop integrals of the gauge field A_μ.
(ii) If there is a single charge at rest $(j_0 \neq 0)$, we see from top equation on the right that the charge gets 'dressed' with a magnetic flux $(F_{12} \neq 0)$, or the other way around a given flux quantum may attract charge and thereby creating a dually charged *anyon*. Integrating the charge distribution one obtains the relation between the flux Φ and charge q of the anyon, $\Phi = 2q/\lambda$. This in turn means that if two of those anyons encircle each other one obtains a phase factor $\exp(-iq^2/\lambda)$, which can take all kinds of values.
(iii) the second equation describes the effect of applying a voltage across the sample; the resulting (Hall) current is perpendicular to the electric field. We see that this Chern-Simons term induces exactly the properties we have described before.

Chern-Simons theory. The Chern-Simons theories are playing a fundamental role in modern physics and mathematics. The American mathematical physicist (and outstanding string theorist) Edward Witten from the Institute for Advanced Study in Princeton, recognized its relevance for three-dimensional topology and the associated physical phenomena. In 1983 he noted that the Chern Simons action provides an intrinsically three-dimensional definition of knot invariants, and as one is free to choose the gauge group, it defines an infinity of them. For this work he was awarded the Fields medal, the mathematical equivalent of a Nobel prize, in 1990. Secondly, Witten showed that if we look at the theory on spaces with a boundary, the theory can be entirely described by an equivalent (1+1)-dimensional conformal invariant field theory on the bound-

ary which is a striking example of the holographic principle we discussed at the end of Chapter I.4 in the context of black holes. Finally, Witten also showed that Einstein's theory of gravity in three space-time dimensions is actually a Chern-Simons theory where the gauge group is the group of local translations and Lorentz transformations. This provides an exciting laboratory to explore the ideas of the holographic principle etc. And as we emphasized in this chapter, topological field theory has become an indispensable tool for the description and understanding of a wide variety of topologically ordered phases in condensed matter.

Topological quantum computation. These topological systems can be characterized by certain symmetries which are quite hidden, and an example of what are nowadays called *quantum groups* or *Hopf algebras*. There is a rapidly growing interest in this field of *topologically ordered media* and more recently also materials called *topological insulators*, which exhibit topological order in three dimensions. These media appear to be quite ideal candidates for quantum information storage and processing, exactly because one can change the state by moving particles around each other. Loosely speaking a computation is nothing but a particular complicated braid or knot of a 'register' of anyons in space-time. It is an *intrinsically fault tolerant* way of doing quantum computations because topological moves are insensitive to local perturbations, that is perturbations caused by local interactions and that is all we have been talking about. No surprise therefore that many think of this as a development of great significance. And by many I not only mean scientists, but also security bosses of public organizations and others who have to hide big $ecret$ behind huge numerical keys which were once believed to be unbreakable, but not in the future. Just wait for quantum technologies to come and get them.

Quantum critical points. Figure III.3.19 shows the phase diagram of what is called a *strange metal*, which is characterized by an anomalous quantum critical phase in which

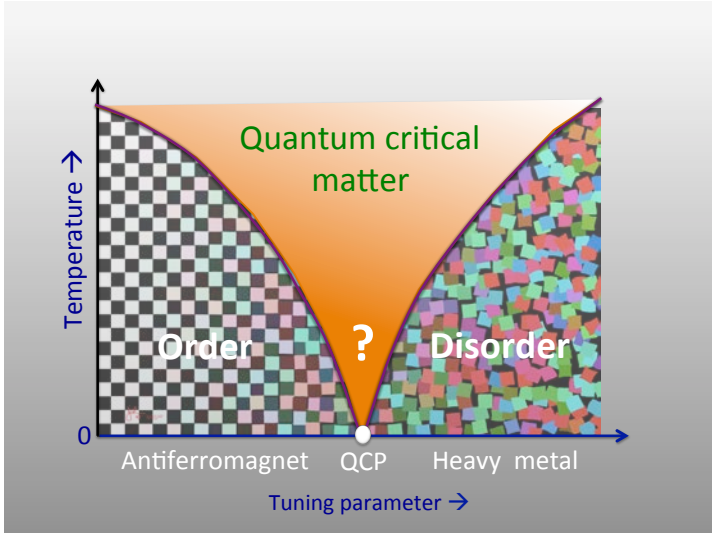


Figure III.3.19: *The quantum critical point.* A quantum critical point separates at zero temperature an ordered (antiferromagnetic) and a disordered phase (a heavy fermion metal). For finite temperature it opens up a region of quantum critical phases, such as what are called 'strange metals.'

the electrical resistivity varies linearly with temperature. This behavior shows up not only at a singular *quantum critical point* (QCP) at zero temperature, but over an extended range of a relevant tuning parameter in the phase diagram. This highly unconventional behavior has defied description within the standard model for metals.

This provides for a new topic that is vigorously pursued at present, and there appear to be a variety of systems that exhibit such a quantum critical point. The general picture that emerges is now that at the quantum critical point, the system can be modelled by an interacting (2+1)-dimensional conformal field theory. This effective theory may, depending on the case, describe emergent Dirac fermions, scalar (Higgs-like) fields and even emergent $\mathrm{U}(1)$ gauge fields. So, in a sense many of the previously known models based on principles of gauge invariance, symmetry breaking and so on make a surprising comeback on a totally different stage. But what is most striking is that the original

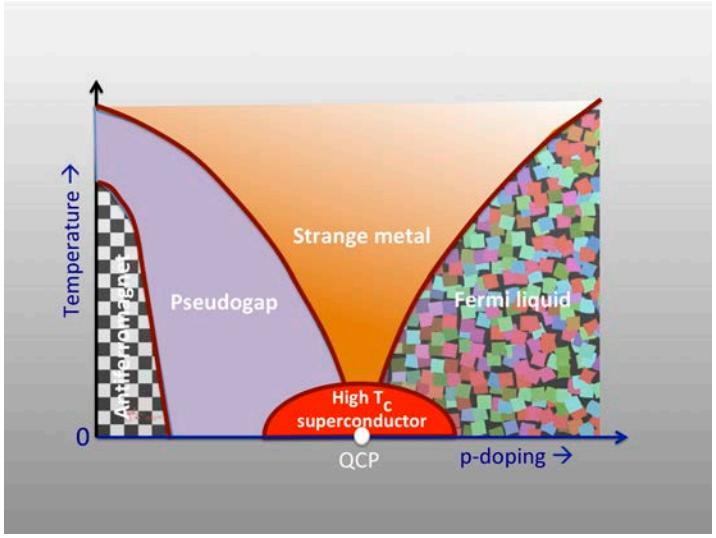


Figure III.3.20: *High* T_C *superconductivity.* The proposed more complicated phase diagram for the *cuprates* exhibiting a high T_C superconducting phase near a quantum critical point (QCP).

electron degrees of freedom are strongly entangled over large distances, and manifest themselves in vastly different guises. One says that these conformal phases are no longer 'adiabatically' connected to the original Fermi liquid phases. There is no smooth way to connect the two regimes.

What makes the quantum critical point relevant is that the behavior persists away from the critical point. So for example there is a well-accepted view that *high* T_C *superconductivity*, which is effectively realized in the two-dimensional layers of certain materials denoted as *cuprates*, is governed by such a QCP as we have indicated in the phase diagram of Figure III.3.20. So the high temperature superconducting phase would be described by a finite temperature version of the (2+1)-dimensional conformal field theory in question. New insights in these theories, which have been inspired by theories of quantum gravity, like string theory and the AdS-CFT correspondence that we discussed before, definitely look promising in a bid to unravel the mysteries of these strange metals. String theory and hard-core condensed matter theory seem strange bed fellows at first sight, but apparently science doesn't know of any taboos in that respect.

Further reading.
On condensed matter physics:

- *Introduction to Solid State Physics*
  Charles Kittel
  Wiley (2004)
- *Solid State Physics*
  Neil W. Ashcroft and N. David Mermin
  Thomson Press (2003)
- *Principles of Condensed Matter Physics*
  P. M. Chaikin and T. C. Lubensky
  Cambridge University Press (1995)
- *Modern Condensed Matter Physics*
  Steven M. Girvin and Kun Yang
  Cambridge University Press (2019)

On superconductivity:

- *Introduction to Superconductivity*
  Michael Tinkham
  Dover Publications (2004)

On topological media:

- *The Quantum Hall Effect*
  Daijiro Yoshioka and D. Yoshioka
  Springer (2010)
- *Introduction to Topological Quantum Computation*
  Jiannis K. Pachos,
  Cambridge University Press, 2012
- *Quantum Phase Transitions*
  Subir Sachdev
  Cambridge University Press (2011)

Chapter III.4

SCALE dependence

In this chapter we explore the notion of scaling*. How does the behavior of physics change if one changes the length or momentum scales?*
We start with some simple geometrical examples of scaling, leading to the notions of scale invariance, self-similarity. and fractals. We move on to discrete maps like conformal mappings used by Escher and dynamical systems like the logistic map.
The next step is to study scaling in physical models, both classical and quantum. This culminates in the notion of renormalization in quantum field theory and the wonderful idea of running coupling constants. We discuss what scaling tells us about the asymptotic behavior of physical theories like the standard model and the possibility of (grand) unification in theories of the fundamental interactions. Finally we point out the profound link between scale (and conformal) invariance and critical behavior

What sets the scale?

When children start building bridges with LEGO they learn what construction engineers know too well: if one simply keeps scaling up the size of a construction it will at a certain point collapse. By simply scaling we mean that we multiply all linear sizes by some given factor. One cannot simply multiply all beam sizes by a factor two to construct a bridge that will span a river twice as wide. The basic reason for this breakdown of scaling was given by Galilei in his discourse on the two world systems, and boils down to the basic observation that the mass of a beam scales as a volume, that is a length cube, while the strength of the beam would only grow with the transverse area meaning a length square. And because the cubic power grows faster than the square, at a certain scale the beam has to break under its own weight.

The question 'what sets the scale' is a vital one, which one had better address before embarking on detailed calculations. In physics the answer is determined by, and expressed in the available dimensionful parameters of the model one employs. Educated guesses are then based on what is called *dimensional analysis* of the parameters that are present in the problem. A given particle mass for example sets a relevant energy scale in a theory in the sense that it separates two regimes defined by energies much smaller and much larger than that mass. One expects that at low energies that mass is so big that the particle will not be excited and therefore will play a negligible role, whereas at high energies the field will effectively behave like a massless field mediating long range interactions, and you expect it to be relevant.

A mass is a dimensionful parameter, and it raises the question what it means to have dimensionless parameters. We have already extensively exploited this principle of dimen-

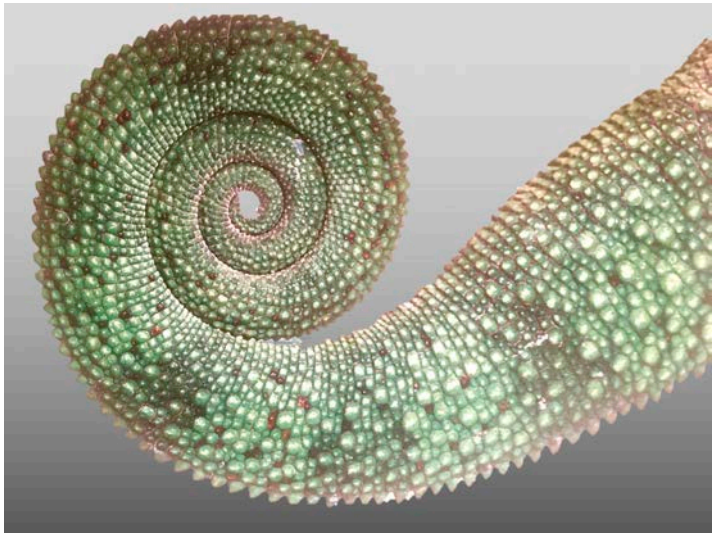


Figure III.4.1: *The spiralling tail.* This is the spiralling tail of the *panther chameleon* living in Madagascar.

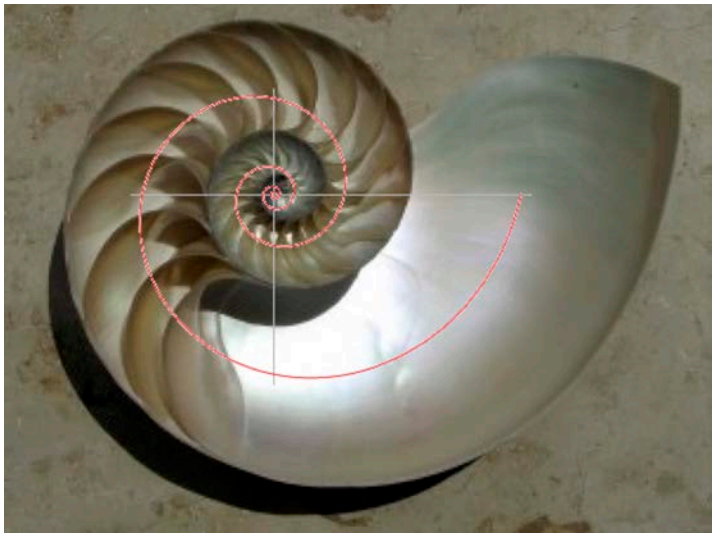


Figure III.4.2: *The spiralling snail house.* This is a beautiful cut of a multi-chamber spiralling house of a snail. The superposed red spiral is a so-called Fibonaci spiral that gives a reasonable approximation.

sional analysis in Chapter I.3, devoted to universal constants, scales and units.[1] At this point one is tempted to claim that a theory with only dimensionless parameters is necessarily invariant under rescaling. In that perspective it furthermore appears that the behavior of a theory at energies much larger than the masses present in the theory will approximate that of some scale invariant model. Interestingly it turns out that this rule of thumb fails in a fundamental way in the quantum domain. This puzzle demands a careful analysis of scale invariance in the quantum domain, a topic that we explore towards the end of this chapter.

We start by showing some relatively easy to envisage geometrical examples of scaling linked to fractals and self-similarity. Next we consider simple dynamical systems where scaling occurs as a function of the parameter in the model. This situation represents a more abstract setting for the property of scaling and (broken) scale invariance. The first is just the *logistic map* an iconic model which exhibits the interesting property of deterministic chaos as the limit of an infinite sequence of period doubling transitions in the space of solutions. Finally, we turn to particle dynamics and field theory both from the classical and quantum point of view. The most surprising and also most difficult to understand results concerning scaling are to be found in quantum field theory and generally in many-particle systems. The crucial observation to be made is that scaling can be interpreted as the model following a calculable trajectory in the parameter space of a class of models. And these trajectories may end on certain fixed points where the theory becomes scale invariant. However, depending on the initial conditions the trajectory may also run off to infinity in which case the theory loses its validity and predictive power. This is usually a call for other may be new physics to be taken into account.

[1] In this chapter we will adopt the natural units $\hbar = c = 1$, (except where explicitly indicated otherwise) which means that we can express all dimensional quantities in units of length, denoted as $[x] \sim \ell$, or in units of mass (or energy) denoted by $[mass] \sim kg$, which scales as inverse length: $[mass] = [length]^{-1} \sim \ell^{-1}$. I will from here on express all quantities in units of length.

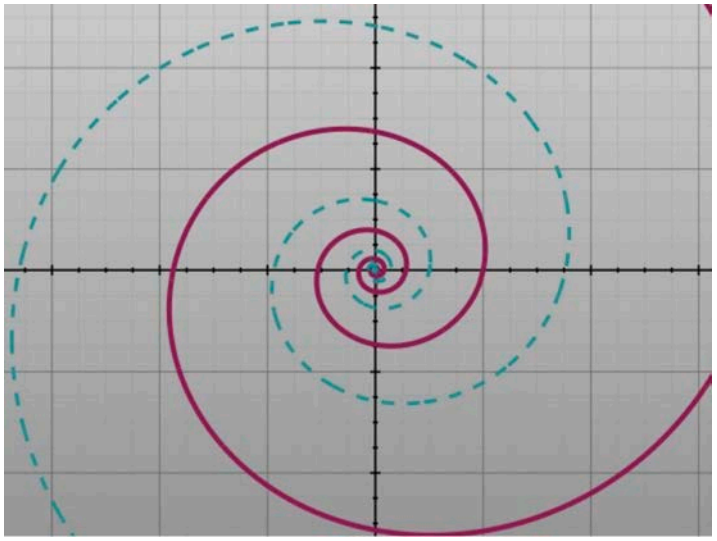


Figure III.4.3: *The logarithmic spiral.* The spiral is given by the equation (III.4.1) corresponding to the red curve. Under a scale transformation $r \to \lambda r$ the spiral is rotated over an angle $\ln\lambda$ corresponding to the blue dashed curve. The curve is therefore strictly invariant only under the discrete set of transformations where $\lambda_n = \exp 2\pi n$.

Scaling in geometry

Self similarity and fractals

Scaling. If we scale an object, we mean to say that under a rescaling of the coordinates it transforms into a larger or smaller conformal object, an object with the same shape but of a different size. If we say that something scales, we mean that it has a specific behavior under scale transformations. For example we may have a geometric object like a triangle and ask how it scales when we divide all coordinates by a factor two, evidently it transforms to a triangle 'half the size'. This means that the lengths of the sides become half as long, and therefore that the area becomes one-fourth the original area. If we say that a property scales, we mean that it scales like a length to some power d, and d is then called its *scaling dimension*. So a 'volume' has a scaling dimension three and a 'point' has scaling dimension zero. This definition basically coincides with what is called the *topological dimension* n of the (vector) space $\mathbb{R}^n$, in which the object is naturally embedded.

So, in this section we address the interesting scaling properties are of certain geometric structures and constructions.

Scale invariance. If the object were to be the real line $\mathbb{R}$, then the scale transformation $x \to x/2$ would map the line on itself, and we therefore say that the line as a whole is *scale invariant*. Similarly the spaces $\mathbf{R}^n$ are scale invariant. So in that sense scale transformations are part of the space-time symmetry like translations, rotations or Lorentz transformations. However the latter do not change the sizes of things, and therefore leave the space-time metric (which defines the notion of distance and therefore size) invariant. As scale transformations affect the size we expect the metric to change by some overall scale or conformal factor.

The logarithmic spiral. A spiral is a wonderful geometric object that has found many stunning applications in nature as an efficient format for growth. We show two examples in Figures III.4.1 and III.4.2. We recommend reading the beautiful chapter on 'The equiangular spiral' in the famous book *On growth and form* of D'arcy Wentworth Thomson, first published in 1942. The 'equiangular spiral' is just the *logarithmic spiral* depicted in Figure III.4.3, and it is specified by giving the polar angle as a function of the radius:

$$\theta(r) = \ln r\,. \tag{III.4.1}$$

Under a scale transformation $r \to \lambda r$ we find that $\theta \to \theta' = \ln\lambda r = \ln r + \ln\lambda$,: in other words we get the same curve back but rotated over an angle $\ln\lambda$. So we could say that it is invariant under a combined scale transformation and rotation over an angle of $\ln\lambda$, or we could say that it is strictly invariant under the discrete subset of scale

transformations, where $\lambda_n = \exp 2\pi n$.

The Cantor set. The Cantor set can be constructed by iterating a map starting by removing the middle third of the closed unit interval $[01]$: in other words $C_1 : [0,1] \to [0,1/3] \cup [2/3,1]$ and the unit interval is mapped to the disjoint union of two smaller copies of itself. The first few iterations of this map are illustrated in Figure III.4.4. If one keeps iterating indefinitely one obtains a tree that is self similar, in the sense that every subtree is identical to a scaled version of the original tree, and one says that this set is *self-similar*. It is the prototype of a *fractal*, which is a term that refers to its dimensionality.

The Hausdorff dimension. A fundamental property characteristic of the scaling property of a fractal is its non-integer *Hausdorff dimension*, which follows from the map that defines the set. At each step we generate a number of copies which we call m, and a factor s by which it is scaled down. For the Cantor set in the figure we have $m = 2$ and $s = 3$. The Hausdorff dimension is defined as $d = \ln m / \ln s$, and for the Cantor set we get the non-integer value $d = \ln 2 / \ln 3 = 0.631$. It is a fractal indeed. The definition recovers the integer topological dimensions for a line, an area or a volume, as that would amount for example to filling a square with four squares of half the size, indeed yielding $d = \ln 4 / \ln 2 = 2$.

Measure zero. The Cantor set itself is a curious mathematical object: it is an infinite set of boundary points of (length) measure zero. If we start with the unit interval of length 1, then at each step we take out $1/3$ of each subset. So the length that is left over after n iterations is $L_n = (2/3)^n$ which tells us that $L_\infty = 0$, showing that it is indeed a set of measure zero.

The Devil's Staircase, Related to this set is Cantor's function depicted in Figure III.4.5. It is a function that maps the unit interval onto itself, but it is not one-to-one. The function is constant on all regions of the interval that are taken out by the infinite iterative process. This function is also called 'The Devil's Staircase' and satisfies an intriguing functional equation:

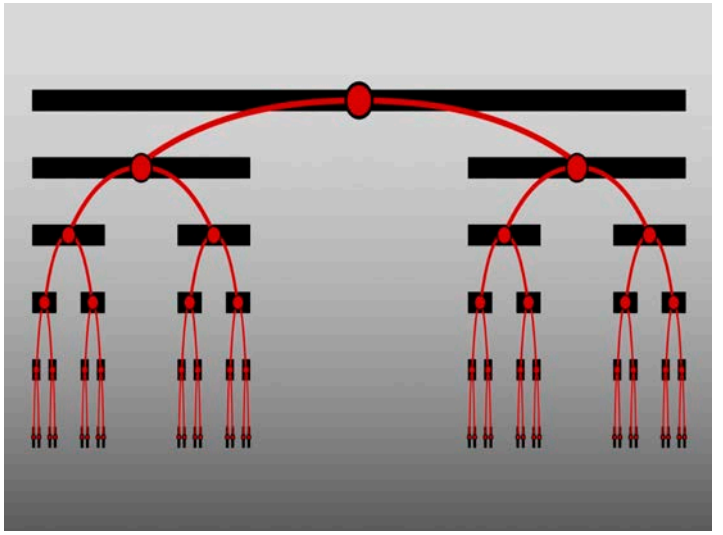


Figure III.4.4: *The Cantor set.* The Cantor set as the result of the infinite iteration of a map where the middle third of the interval is removed starting with the closed unit interval $[0,1]$. The resulting set is the prototype of a *fractal (string)*, clearly displaying the property of *self-similarity*. (Source: Sam Derbyshire)

$$f(x) = 2f(\frac{x}{3}) \quad x \in [0,1]\,, \tag{III.4.2}$$

that fully captures its scaling behavior. The equation says that if we first cut off the curve at x=1/3 and scale it up horizontally by a factor three, and after that vertically scale it up by a factor two, we get the original function back. This formula encapsulates its scale invariance property. An instructive way to think about this function is to look at it as the $n \to \infty$ limit of an iterative approximation scheme defined by:

$$f_n(x) = 2f_{n+1}(\frac{x}{3})\,,$$

with initial condition $f_0(x) = x$. So indeed, this staircase is develish in that it has an infinite number of steps that in some regions become extremely narrow.

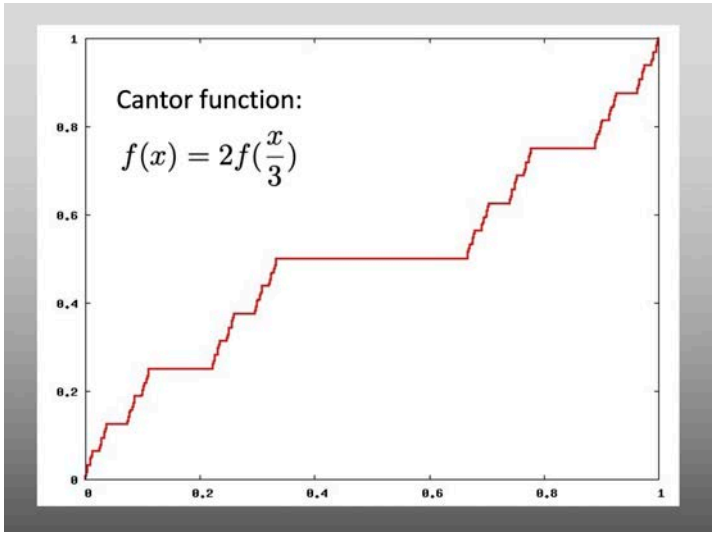


Figure III.4.5: *The Devil's staircase.* An alternative way to represent the Cantor set is as a function from the unit interval $[0,1]$ onto itself, by Cantor's function, also known as the Devil's Staircase. It is constant on the sub-intervals taken out, and has a constant slope in between. One can guess the scaling property of this function from looking at it: it satisfies the functional equation $\mathrm{f(x)} = 2\mathrm{f(x/3)}$, which captures the self similarity of the function.

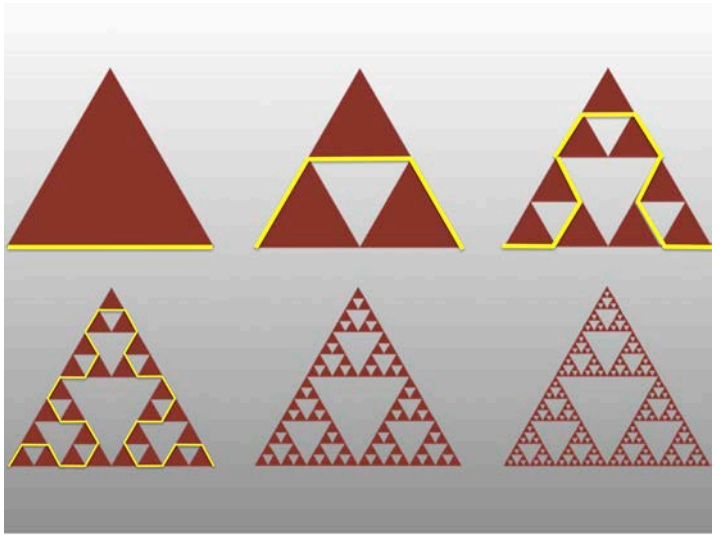


Figure III.4.6: *Sierpinski gasket.* This geometrical structure has fractal properties. It is a self-similar structure. If we take the number of scaling steps to infinity it becomes fractal. If we scale the dimensions by a factor 2, then the length of the yellow curve does not increase by a factor 2 but by a factor 3. This means that the scaling dimension of the gasket would be $\mathrm{d} = \ln 3/\ln 2 = 1.58$

The Sierpinski gasket. A slightly a more complicated example is the *Sierpinski triangle* or *gasket* of Figure III.4.6, which is obtained by iterating a discrete map of a shape in to a scaled version of itself. It generates an object which is self-similar by construction. And if we iterate the mapping indefinitely we would end up with a fractal space that would be invariant under a specific set of discrete scale transformations.

The Hausdorff dimension involves again a length downscaling factor s, which for the Sierpinski triangle equals $s = 2$, and a multiplication factor $\mathrm{m} = 3$ as is clear from the figure. Therefore the gasket has the *fractal* dimension: $\mathrm{d} = \ln 3/\ln 2 = 1.58$.

In the figure we have also drawn a yellow fractal curve and we may apply the same argument, and because for a line segment we have again $s = 2$ and $\mathrm{m} = 3$ we find the same value for the fractal dimension, $\mathrm{d} = 1.58$, validating our intuition that the dimension of the gasket is more than one and less then two. We may also look at the measure of the objects, the area covered by the purple triangles after k iterations equals $\mathrm{A_n} = (\frac{3}{4})^{\mathrm{n}}\mathrm{A_0}$, which means that the limiting area would be $\mathrm{A}_\infty = 0$, so we find again a set of measure zero. The length of the fractal curve would tend to infinity and its measure is unbounded.

The disc where Escher and Poincaré met

In Figure III.4.7 we depicted a sequence of images that interpolate smoothly between the original Escher art work *Circle Limit II* and its underlying hyperbolic geometry of the disc. This hyperbolic tessellation (or tiling) is composed of

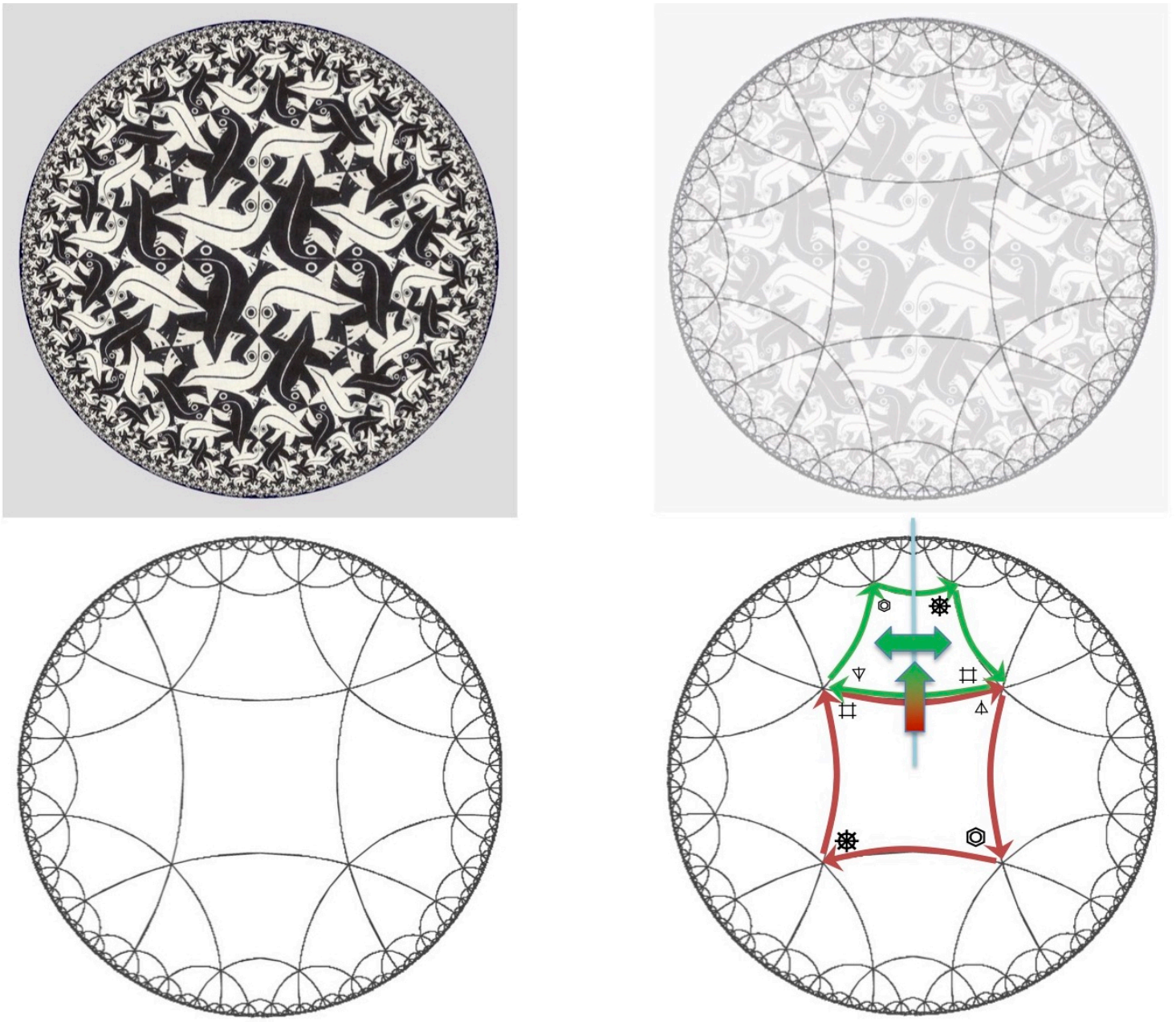


Figure III.4.7: *The hidden geometry of Escher.*

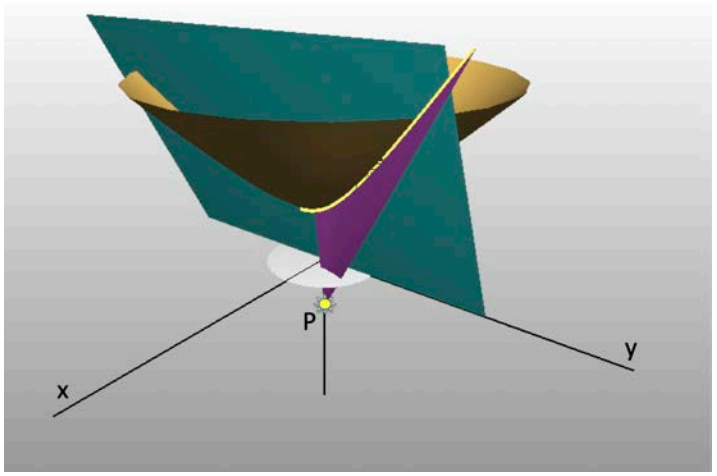


Figure III.4.8: *Poincaré disc.* In the figure we show how to get from the hyperbolic plane in orange with geodesics that are hyperbola like the yellow one. These are obtained by intersecting with a plane through the origin like the green one. The disk obtains by stereo-graphically projecting the hyperbolic plane down to the unit disc in the $z = 0$ plane to the point $P = (0, 0, 1)$.

circular segments that intersect the unit circle orthogonally. Starting with the hyperbolic square at the center one obtains the subsequent segments or vertices by mirroring (inverting) the points in the various circular segments as we indicated in Figure III.4.9. The radial tree connecting the nodes is very much like the binary tree used to construct the Cantor set as displayed in Figure III.4.4.

The hyperbolic plane. For the hyperbolic plane we may choose the positive $z > 0$ sheet satisfying the equation $x^2 + y^2 - z^2 = 1$. It is the yellow surface in Figure III.4.8. This hyperbolic plane is not so unfamiliar as you might have thought; it is identical to the plane defined by the relativistic energy-momentum vectors p_μ for a particle with unit rest-mass living in a flat two-plus-one–dimensional Minkowski space-time which we discussed in Chapter I.1. You can also view it as the Minkowskian analogue of the unit sphere in three Euclidean dimensions (or rather the Northern hemisphere thereof), which obtains if one switches the sign in front of the z^2 term. The geodesics on the hyperbolic plane correspond to any intersection of the surface with a plane through the origin like the green plane in the figure yielding the yellow hyperbola. These hyperbolas are geodesics to be compared with straight lines on the plane or the great circles on an ordinary spherical surface.

The Poincaré disc. The disc geometry that Escher exploited corresponds to the so-called *Poincaré disc*, which is the stereographic projection of the hyperbolic plane on the unit disc in the flat z=0 plane (light grey in the figure) from the point $P = (0, 0, -1)$. For a given hyperbola one gets a line bundle like the purple surface in the figure, yielding a circular segment that approaches the circle bounding the disc orthogonally as indicated in Figure III.4.7. This bounding circle represents the circle at infinity on the hyperbolic plane. These segments accumulate towards the boundary circle which represents a critical point, or a limit like we described in the previous examples. A wonderful non-Euclidean construction indeed.

The Escher tilings. That fractal geometry of hyperbolic tessellation of the disc clearly exhibits how the basic 'amphibian' gets rescaled and rotated if one approaches the boundary, and indeed the number of them tends to infinity near the boundary. The different hyperbolic tilings can be denoted by a pair of integers $\{n, k\}$, called a Schäfli pair, where n is the number of edges of the basic polygon ($n = 4$ in this case), and k is the number of edges that meet at a vertex ($k = 6$) under equal angles, equaling $360/k$ degrees. Clearly the n angles of the polygon add up to $360\,n/k$ degrees, and if this sum is less than 360^o, then we are dealing with a regular tiling of the hyperbolic plane. Note that in Chapter III.2 in the section on crystal structures we discussed the tilings by regular polygons of the plane, where indeed the condition $k = n$ could only be satisfied for $k = 3$, 4, and 6.

Maurits Escher himself was not a mathematician, but his

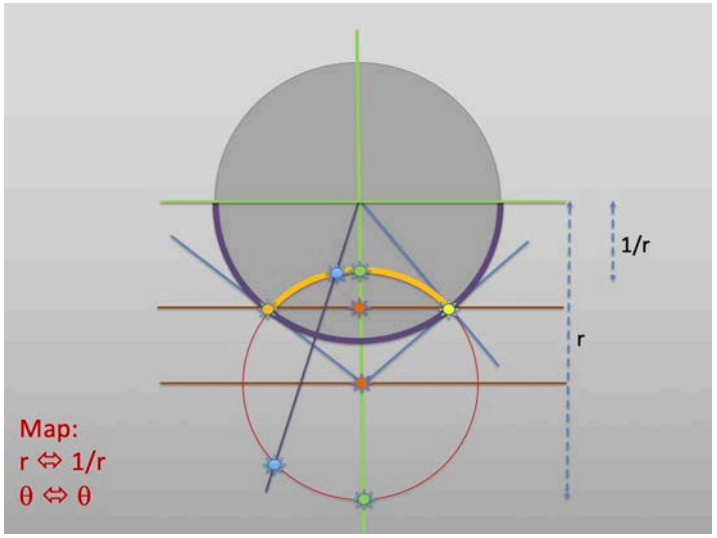


Figure III.4.9: *Inversion map.* This map defines for any point (r, θ) outside the unit circle bounding the disc a mirror point $(1/r, \theta)$.. If a circle crosses the disc, points on the inner and outer segments connected by a radial line through the center are mapped onto each other. The Escher disk combines this mirroring in ever smaller circles with mirroring in a symmetry axis through the center of the disc, as is indicated in the last picture of the previous figure.

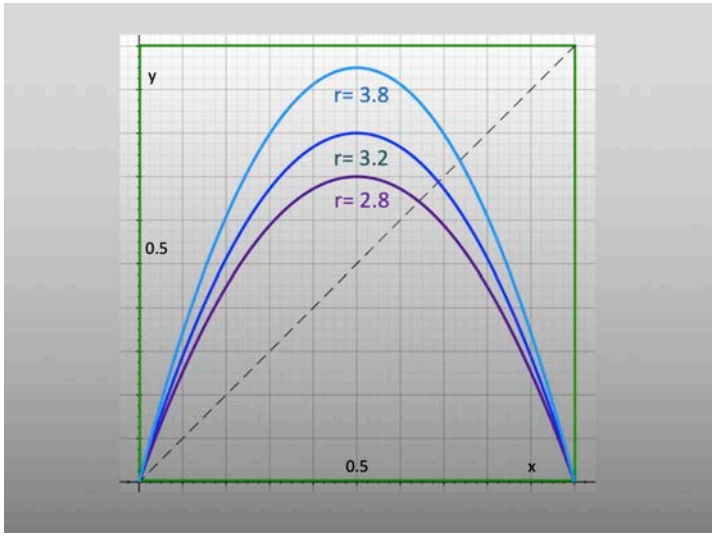


Figure III.4.10: *Logistic map.* This iterative map defines a discrete dynamical system on the unit interval $(0 \leqslant x \leqslant 1)$ and is given by $x_{n+1} = f(x_n) = r\, x_n(x_n - 1)$.

work - not surprisingly - attracted much attention from mathematicians. This started at the International Congress of Mathematicians in Amsterdam in 1954, where one of the organisers, N.G. de Bruijn, had arranged for an exhibition of Escher's work in the *Stedelijk Museum*.[2] In particular the British mathematician H.S.M. Coxeter had many exchanges with Escher on the mathematical meaning and interpretations of his work. It is clear that the interactions fascinated and inspired Escher, but it is also clear that he kept doing the mathematics in 'his own way:'

> My great enthusiasm for this sort of picture and my tenacity in pursuing the study will perhaps lead to a satisfactory solution in the end. ... it seems to be very difficult for Coxeter to write intelligibly to a layman. Finally, no matter how difficult it is, I feel all the more satisfaction from solving a problem like this in my own bumbling fashion.
>
> Escher in a letter to his son George

Escher used the term *coxeteering* for his incredibly imaginative and creative explorations of the hyperbolic disc and its tessalations in a series of prints he called *Circle Limits*.

Scaling in dynamical systems

The systems we have been looking at so far have been completely geometric where the scaling patterns were quite obvious from the start, but now we want to explore the domain of dynamical systems where scaling behavior can be more hidden but highly non-trivial. We start with the *lo-*

[2] For the mathematics of Escher's work I refer to the book edited by H. F. M. Coxeter, M. Emmer, R. Penrose and M. L. Teuber (M.C. Escher: Art and Science) and an article by Doris Schattschneider (Notices of the AMS, Volume 57, Number 6, 2010).

k	$cycle(2^k)$	r_k
1	2	3
2	4	3.449490
3	8	3.544090
4	16	3.564407
5	32	3.568750
6	64	3.56969
7	128	3.56989
8	256	3.569934
9	512	3.569943
10	1024	3.5699451
11	2048	3.569945557
∞	accumulation point	3.569945672

Table III.4.1: The bifurcation sequence.

gistic map which is a (discrete) dynamical system which exhibits scaling behavior in its parameter space $\{r\}$.

The logistic map

The logistic map is a canonical example of a system which displays what is called *deterministic chaos*. It is an iterative map of the unit interval $(0 \leq x \leq 1)$ onto itself, where each iteration corresponds to a time step. The map is quadratic and given by

$$x_{n+1} = f(x_n) = r\, x_n(x_n - 1) \quad (n = 1, 2, 3, \ldots)\,. \quad \text{(III.4.3)}$$

It is plotted in Figure III.4.10 for three different values of the parameter r. This is one of the most well-studied equations in mathematical physics with a vast literature dedicated to its remarkable properties.

In Figure III.4.11 we have in the left column depicted the orbits corresponding to the first fifty iterations of the map with initial value $x_0 = 0.2$, for three values of r. What we see is that with increasing values of r the behavior of the orbit for $n \gg 1$ changes drastically.

For small r it starts with a fixed point, then we get into a region where the orbit becomes a 2-cycle, after which one obtains ever smaller regions where the period doubles to some 2^k-cycle. In the second column the same orbits are represented as a *cobweb diagram* where the successive steps are obtained by mirroring the outcome of the $n-$th iteration in the line $y = x$ to obtain the input for the $(n+1)-$th iteration. In these diagrams the limit cycle behavior is very clear. In the right column we have depicted the so-called *bifurcation diagram*, which shows what the cycles are as a function of r and at what values the period doubling occurs. For increasing r the points r_k, where the period doubles occurs and the 2^k-cycle starts, accumulate at some critical point $r_\infty = 3.56995\ldots$, where a transition to chaotic behavior occurs.

The bifurcation diagram of Figure III.4.12 suggest that there is some form of self-similarity present in this system and it was Mitchell J. Feigenbaum who in 1978 extracted two fundamental constants from the system that characterize the scale invariance of the system near the critical point r_∞.

The first Feigenbaum constant is given by the limiting behavior of the following sequence (see figure):

$$\lim_{k\to\infty} \frac{r_k - r_{k-1}}{r_{k+1} - r_k} = \lim_{k\to\infty} \frac{d_k}{d_{k+1}} = \delta = 4.6692\ldots\,. \quad \text{(III.4.4)}$$

This number δ is universal in that it does not depend on the details of the map as long as it has quadratic behavior near the maximum and vanishes at the endpoints of the interval, and it turns out that this constant governs the asymptotic behavior of all period doubling sequences. One might rephrase the above equation by saying that for large $k \gg 1$ the interval $d_k^* = r_\infty - r_k$ converges like a geometric series $d_k^* \simeq C\delta^{-k}$.

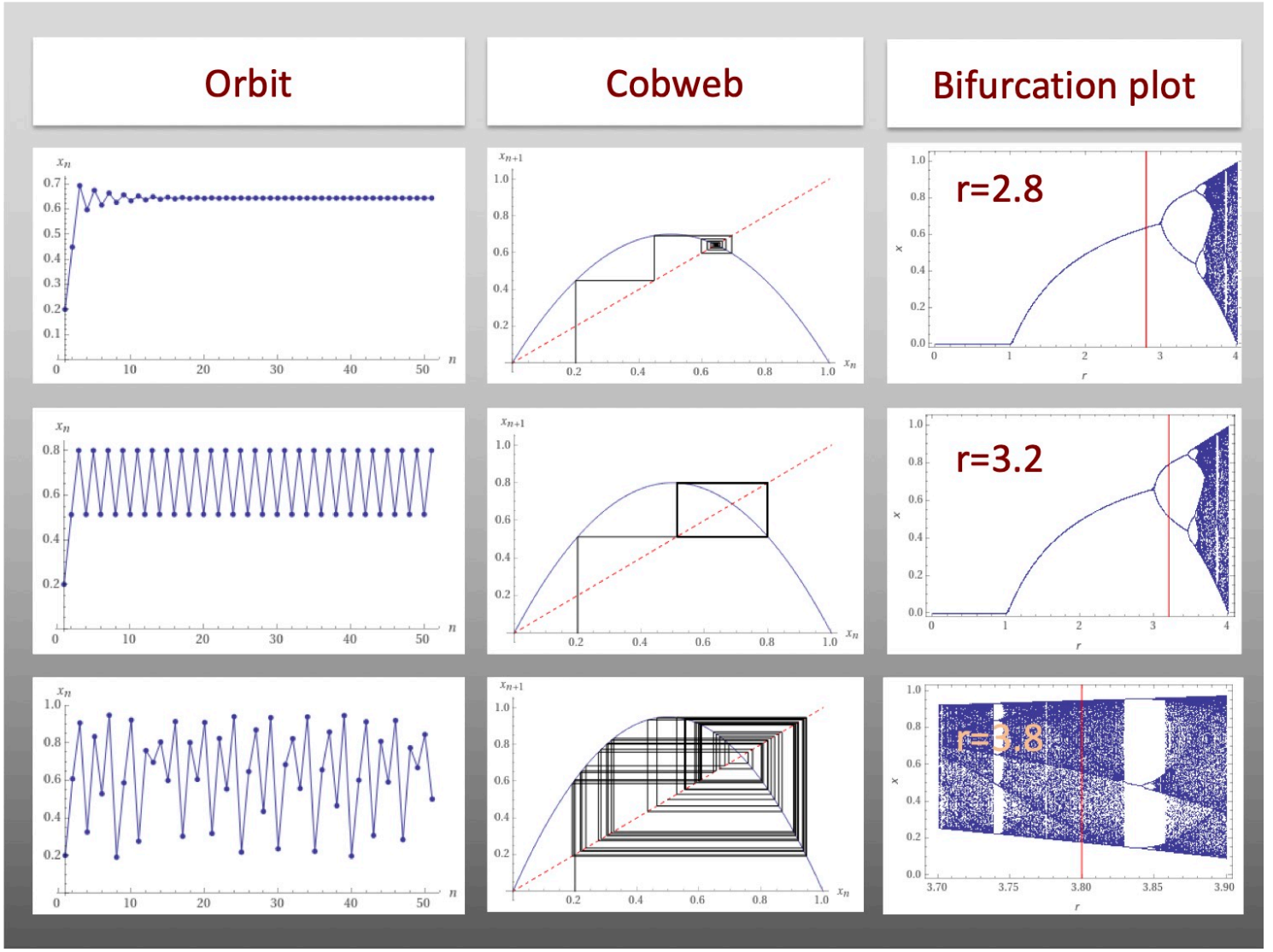


Figure III.4.11: *Logistic map orbits.* We show orbits starting at $x = 0.2$ for different three different r values ($r = 2.8, 3.2$ and 3.8) in the first column. In the second column the same orbits are given as 'cobwebs.' In the final column we marked the corresponding r values in the bifurcation diagram.

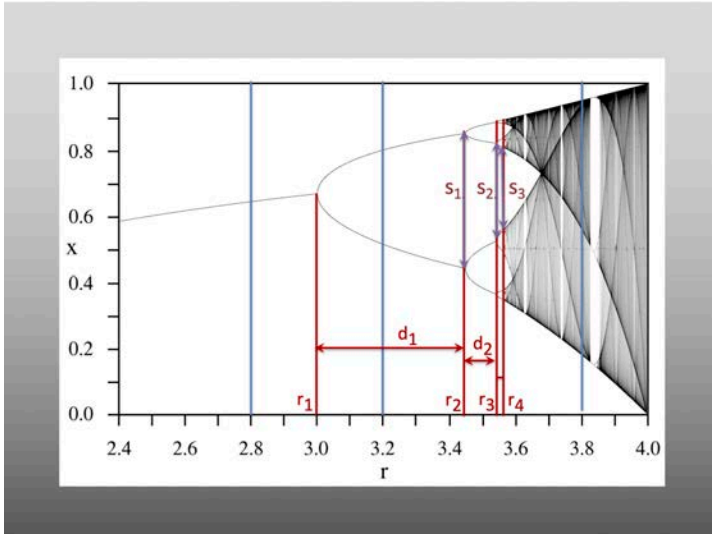


Figure III.4.12: *Bifurcation diagram.* This diagram gives the x values in the subsequent 2^k limit cycles as a function of the r parameter of the logistic map. The limiting behavior does not depend on the initial value x_0, and forms therefore a *global attractor*. Starting at small r the sequence of points r_k where the doubling to a 2^k-cycle starts accumulates at some point $r_\infty = 3.56995...$, after which a highly unpredictable limiting behavior sets in, which is called *deterministic chaos.*

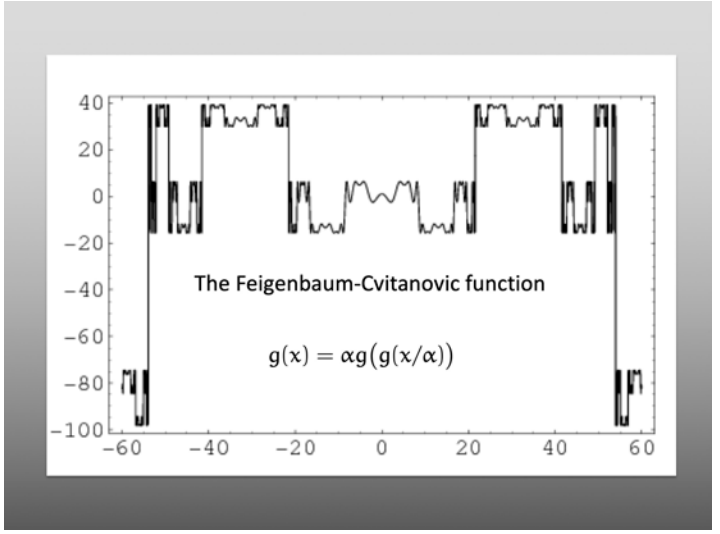


Figure III.4.13: *Feigenbaum-Cvitanovic function.* The F-C function can be compared with Cantor's staircase function. It captures the strange attractor of the logistic map. The function satisfies the F-C functional equation $g(x) = \alpha g\big(g(x/\alpha)\big)$.

There is a second universal constant that can be extracted from the diagram. It is determined by the limiting behavior of the sequence of separations s_k, where s_k is the separation in x between the two adjacent central values of the 2^k-cycle at $r = r_{k+1}$, as we have indicated in the figure. For large k one finds that $s_{k+1} = s_k/\alpha$ where $\alpha = 2.5029\ldots$.

The essential scaling property of the limiting behavior of the period doubling sequence is expressed by a scaling function $g(x)$, which would be the solution of a functional[3] equation analogous to equation (III.4.2) for the devil's staircase. The equation for the period doubling sequence is called the Feigenbaum-Cvitanovic equation:

$$g(x) = \alpha g\big(g(x/\alpha)\big)\,, \tag{III.4.5}$$

with boundary condition $g(0) = 1$. There is a unique solution to this equation that fixes both the value of α and the function $g(x)$ which you should think of as specifying the attractor at the accumulation point (the set of 2^k points in the limit $k \to \infty$). The F-C function is plotted in Figure III.4.13 and could have been called the 'devil's castle' because the embattlements contain ever smaller self-similar versions of the castle. A stunning architectural masterpiece obviously. A remarkable property of this equation and thus its solution is that it is independent of the precise form of the logistic map f, and it is in that sense that the parameter alpha is *universal* over the class of functions denoted by $\{f\}$.

[3] A *function* $f(x)$ is a mapping from a space $\mathcal{X}$ of the variable to some space of function values, like the real line $\mathbb{R}$ or the complex plane $\mathbb{C}$. Formally a *functional* is a 'function of a function' and corresponds to a map from a space or a certain class of functions $\mathcal{F}$ to a space of values like $\mathbb{R}$ or $\mathbb{C}$.

Scaling in quantum theory

Quantum mechanics

In earlier chapters we have argued that (continuous) space-time symmetries lead to conserved quantities, and that in quantum theory these conserved quantities are represented by certain operator expressions that act on the Hilbert space. These operators are expressed as functions of the basic degrees of freedom. So, in the quantum mechanics of a particle the basic operators are X and P, corresponding to the classical phase space coordinates x and p. And from these one can construct the operators for other dynamical variables like the energy or angular momentum. The operators work on the Hilbert space of wave functions. In quantum field theory the basic operators are the fields themselves and their conjugate momentum fields and these work on the multi-particle Hilbert space.

Operators that represent space-time symmetries. In a quantum system symmetry operators commute with the Hamiltonian, and therefore transform states that have the same energy among each other: in other words, states that are degenerate. We recall that for the case of the hydrogen atom, the energy levels are labeled by the principal quantum number n, and for any n we have an n^2 degenerate set of states. This set consists of representations of the rotation group $SO(3)$ labeled by the angular momentum eigenvalues l, with $l = 0, \ldots, n-1$. At a given energy level n the total degeneracy can be understood if one adds the Runge-Lenz vector, to be thought of as a vector of symmetry operators to the symmetry algebra. This is a dynamical symmetry which follows from the particular form of the Coulomb (or Newton) potential and is not related to an underlying space-time symmetry. Inclusion of this vector extends the symmetry algebra from $so(3)$ to $so(4)$, as we discussed in connection with Figure II.6.3 in Chapter II.6.

Let us now turn to the expression for the operator Λ that generates scale transformations on a one-particle Hilbert space. We do so after we have recalled how it worked for the case of translations.

The case of translations generated by momentum. In previous chapters we discussed how in quantum theory the momentum operator P acting on a wave function is represented as the Hermitean differential operator $P = -id/dx$ ($\hbar = 1$). This operator generates 'translations' meaning to say that if we act with a finite transformation on any function

$$T(a)f(x) \equiv e^{iaP} f(x) = f(x+a)\,,$$

then the argument of the function is shifted by an amount a. The momentum operator has a continuous set of eigenfunctions $f_k(x) \simeq e^{ikx}$ because:

$$P\, f_k(x) = k f_k(x)\,.$$

These functions are periodic and the expansion of an arbitrary function in this basis of eigenfunctions amounts to a Fourier decomposition of that function. Needless to say that the only translation invariant function is the constant function, corresponding to $k = 0$. Finally, we recall that translational invariance of a system implied that the momentum operator would commute with the Hamiltonian, and henceforth momentum would be conserved.

The scaling operator Λ and its eigenfunctions. Now we ask the same questions about scale invariance: what is the operator representing scale transformations on functions, and what are its eigenfunctions, and finally, what does it mean to say that a system is scale invariant? The scale operator is $\Lambda(x) \equiv x\dfrac{d}{dx}$ and its eigenfunctions are quite simple to derive:

$$\begin{aligned} x\frac{d}{dx}\, g_d(x) &= d\, g_d(x) \\ \Rightarrow\ \frac{dg_d}{g_d} &= \frac{d}{x} \Rightarrow \ln g_d = d \ln x = \ln x^d\,. \end{aligned} \tag{III.4.6}$$

So again there is a continuum of eigenfunctions which are just powers of x : $g_d(x) \sim x^d$ for any d. The eigenvalue d is called the scaling dimension. Under a finite scaling transformation $S(\alpha)$ we would get:

$$S(\alpha)g_d(x) \equiv e^{\alpha x(d/dx)}g_d(x) = e^{\alpha d}g_d(x)\,.$$

This expression gains transparency and elegance if we take the parameter logarithmic:

$$S(\ln\lambda)g_d(x) = e^{d\ln\lambda}g_d(x) = \lambda^d g_d(x) = g_d(\lambda x)\,.$$

Power laws. This gives an alternative way to define a *scaling function* in general; it is any function $h(x)$ that satisfies the scaling law:

$$h(\lambda x) = \lambda^d h(x)\,, \qquad \text{(III.4.7)}$$

for any λ, where the power d is defined as the scaling dimension of the function. Indeed, the scaling functions are the eigenfunctions of the scaling operator and are just single powers of their argument. A scale invariant function is the eigenfunction with $d = 0$, again meaning any constant function.

We just saw that making the scale transformation $S(\ln\lambda)$ on an eigenfunction effectively multiplies the argument of that function with λ. This is a special property in the sense that it multiplies the argument and not the function. Thus, if I apply the operator to an arbitrary linear combination of eigenfunctions, I get exactly the same combination back with scaled argument. In other words, if we think of an arbitrary function that can formally be expanded in a power series, then what the scale transformation S does is just to scale the argument of that arbitrary function. This is to be expected because it is the defining property of a scaling transformation on any function, but it does not imply that any arbitrary function is a scaling function, as it will in general not satisfy the scaling property (III.4.7).

The symmetry algebra including scaling. To further discuss scaling properties it is useful to study its commutation relations with other elementary operators forming the dynamical Lie algebra. For example from

$$[\Lambda, X] = \frac{i}{\hbar}[XP, X] = \frac{i}{\hbar}(XPX - XXP) = \frac{i}{\hbar}X[P, X] = X$$
$$[\Lambda, P] = \frac{i}{\hbar}[XP, P] = \frac{i}{\hbar}(XPP - PXP) = \frac{i}{\hbar}[X, P]P = -P \qquad \text{(III.4.8)}$$

It gives the operator back multiplied by its naive scaling dimension, which is the dimension of the operator in units of length. Note that the angular momentum operator has scaling dimension zero as it involves products of X and P components; this is also consistent with its quantization in integer multiples of $\hbar$ which at this point is dimensionless as it has units $Js \sim \ell^0$.

The calculation we just did shows that we can extend the combined Lorentz and translation symmetry, denoted as the *Poincaré group*, with the scale transformations. Including the scale transformations we also need to include the so-called *inversion operator* I with $I : \mathbf{x} \to \mathbf{x}/x^2$. Adding these two operators to the dynamical operator algebra, one ends up with a closed Lie algebra with fifteen generators, which is referred to as the *conformal algebra* which for four-dimensional Minkowski space is the algebra $so(4,2)$. This algebra corresponds (is isomorphic) to the 'rotations' in a six-dimensional 'space' with four space and two time dimensions.

So far we have mainly discussed mathematical features of scaling functions and operators. Let us now return to the physics of scale invariance. We do this at various levels of increasing complexity starting with simple classical systems and moving up to applications of scaling in quantum (field) theory.

Scaling properties of some Hamiltonians. Having the scale operator it is interesting to see what one can learn about the scaling properties of some Hamiltonians and other operators.

To keep it simple we look at a particle with Hamiltonian $H = U + V$ or Lagrangian $L = U - V$, where kinetic term $U = P^2/2M$ and for the potential we choose a simple power, $V = a_k x^k$. Now for consistency we must have that $[H] = [U] = [V] = \ell^{-1}$. This implies that indeed $[U] = [M]^{-1} \cdot 2[P] = \ell^{-1}$, as expected. For the potential term we find that $[V] = [a_k] \cdot \ell^k = \ell^{-1}$, from which we conclude that $[a_k] = \ell^{-1-k}$, so this simple power counting yields the dimensionality of the parameters or coupling constants.

We see that the kinetic and potential terms will in general scale differently under scale transformations of the coordinates. Just transforming coordinates and keeping the parameters fixed we get that:

$$x \to x' = \lambda x \Rightarrow H \to H' = \frac{1}{\lambda} H(\lambda) = \frac{p^2}{\lambda^2 M} + a_k \lambda^k x^k \,.$$

This expression leads us to conclude that under a rescaling of the coordinates the Hamiltonian is mapped into a similar Hamiltonian $H(\lambda)$, with different, scale dependent, parameters: $M' = M(\lambda) = \lambda M$ and $a'_k = a_k(\lambda) = \lambda^{k+1} a_k$.

Let us look at some simple cases:

1. The harmonic oscillator.
The potential is given by $V(x) = \frac{1}{2} K x^2$, and corresponds to the case $k = 2$. The spectrum is depicted in Figure II.5.14 on page 396 of Part II. It is equally spaced, with energy levels $E_n = \omega(n + \frac{1}{2})$ where the frequency ω is given by $\omega = \sqrt{K/M}$. The frequency is the only physically relevant parameter and we see that its scale dependence is : $\omega(\lambda) = \sqrt{K(\lambda)/M(\lambda)} = \lambda\omega$. The spectrum apparently scales linearly with λ.

The concept that we want to emphasize is the fact that under scaling the theory changes. If we define the theory as a point in the space of parameters, then under rescaling the theory will trace out a trajectory in that space. In the example at hand, the parameter space is a plane with coordinates M and K. With $M(\lambda) = \lambda M$ and $K(\lambda) = \lambda^3 K$, we see that we can eliminate the λ, to obtain a function $K(\lambda) = (K/M^3)M(\lambda)^3$. We have depicted one such trajectory for a particular initial condition $K(1) = K$ and $M(1) = M$ in Figure III.4.14.

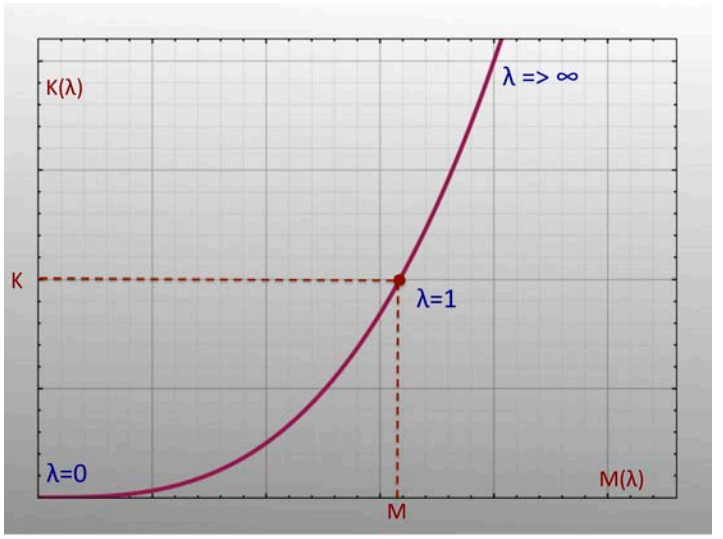


Figure III.4.14: *Scaling trajectory of harmonic oscillator.* Scaling the coordinate by a factor λ in the harmonic oscillator Hamiltonian is equivalent to a trajectory of the parameters $M(\lambda)$ and $K(\lambda)$ through parameter space.

What we learn from this graph is not earth-shattering, just that for large values of λ, the potential term starts to dominate so that the system will get locked into the ground state. On the other hand, if $\lambda \to 0$ the kinetic term dominates and the hamiltonian approaches that of a free particle, where the energy gap tends to zero. So, at short distances the theory has a fixed point where the theory is free, a primitive precursor of the notion of what is called 'asymptotic freedom'. This is not so surprising, because it is what we could have concluded directly from the linear λ dependence of $\omega(\lambda)$, which implies that the energy gap tends to zero.

2. The hydrogen atom.

The part of the Hamiltonian that is of interest here is the radial part because when we scale the coordinates we rescale the r variable and not the angular variables θ and φ. This reflects the fact that the angular derivatives of the Hamiltonian are all contained in the the term $\mathbf{L}^2/2Mr^2$, where the angular momentum operator $\mathbf{L} = \mathbf{X} \times \mathbf{P}$. In view of the relations (III.4.8) the scaling dimension of $\mathbf{L}$ is $d_L = 0$, and thus, as stated before, the angular momentum is scale invariant. So, we are left with the 'radial' Hamiltonian, which is very similar to the one given in equation (I.4.1) we discussed in Chapter I.4, it takes the form:

$$H = \frac{p_r^2}{2M} + \frac{l(l+1)}{2Mr^2} - \frac{e^2}{4\pi r}\,,$$

where l is the angular momentum label, and $l(l+1)$ is the eigenvalue of the operator $\mathbf{L}^2$. Doing the scaling exercise as before we find that $M(\lambda) = \lambda M$ and, interestingly, that the charge does *not* rescale $e(\lambda) = e$. Let us look what that implies for the spectrum in this case, the discrete bound state energy levels are labeled by the principal quantum number n, and are given by:

$$E_n = \frac{E_1}{n^2} \quad with \quad E_1 = M\big(\frac{e^2}{4\pi}\big)^2 .$$

We conclude that the levels simply scale like $E_n(\lambda) = \lambda E_n$, confirming our naive expectations.

On the one-particle level the quantum analysis of scaling properties does not lead to surprising new insights. It merely confirms the behavior you would expect based on naive dimensional analysis. As we will see in the remaining sections of this chapter it is in quantum field theory that interesting complications arise.

Quantum field theory

In this subsection we turn to the question what scaling means in quantum field theory. We will look at this problem from a rather general and abstract point of view, avoiding as many technicalities as possible. In later sections we give more details about how these results can be obtained.

The fundamental question is again to understand how parameters of the model change depending on the scale at which one looks at the system. And as the quantum uncertainty relations imply an inverse relation between spatial scale (wavelength) and momentum or energy, we expect to learn something about the energy dependence of the phenomena the theory describes. By exploiting arguments like the ones we used in the previous subsection we may even probe the domain of validity of certain theories.

Actions and Lagrangians. In general a theory can be defined by its energy function or Hamiltonian H, or its action S. As mentioned before, in relativistic systems and field theories, one prefers the action because it is a manifestly Lorentz invariant quantity, while the energy is not as it is a component of the energy-momentum four vector.

The action can be written as a functional of the field, a space-time integral over a *Lagrange density* $\mathcal{L}$, which is an expression in the fields and their derivatives. We write:

$$S = \int \mathcal{L}\, d^4x\,, \tag{III.4.9}$$

and in units where $\hbar = c = 1$ the action is a dimensionless quantity. At this point the difference between the quantum and classical expression resides completely in the interpretation of the fields. Classical fields are just scalar, or vector, or spinor valued functions on coordinate space. Quantum fields are very different types of objects: they are operator valued and work on some multi-particle Hilbert space as we discussed in Chapters I.4 and II.5.

Three examples. In the remaining sections of this chapter we will refer to the three different examples of Lagrangian densities we introduce next.

- *The ϕ^4 model.* The first action is about the simplest non-trivial field theory one can think of and it owes its popularity exactly to the fact that it is often used to demonstrate the intricacies of quantum field theory. It is a theory of a real scalar field $\phi(x^\mu)$ with a quartic self-interaction. The action of this so-called 'ϕ-fourth theory' is defined by the relativistic Lagrangian density $\mathcal{L}$:

$$\mathcal{L}(\phi, \partial_\mu\phi) = \frac{1}{2}(\partial_\mu\phi)^2 + \frac{1}{2}m^2\phi^2 + \frac{\lambda_4}{4!}\phi^4\,. \qquad \text{(III.4.10)}$$

The classical field ϕ is just an arbitrary function which may be expanded in an orthonormal set of basis functions, for example energy momentum eigenstates or planes waves $\{\phi(k)\}$:

$$\phi(x) \sim \int \phi(k)\, e^{-ik\cdot x}\, d^4k\,.$$

In Chapter II.5 we pointed out that in quantum field theory the fields are operators acting on a multi-particle Hilbert space and can create or annihilate particles in any given energy momentum state labeled by k^μ. with $k^2 = m^2$ (m = rest mass). The first two terms of the Lagrangian are often denoted as $\mathcal{L}_0$, and being quadratic in the fields, they make up the free field theory. The last term denoted by $\mathcal{L}_{\rm int}$ describes the self-interactions of the field with coupling strengh λ_4.

- *The toy model.* Of course a field theory can be defined in any number of space-time dimensions, and formally nothing forbids us, for pedagogical reasons, to restrict ourselves to a theory with only a time dimension. Then the field becomes just like a time-dependent position coordinate $\phi(t) \sim x(t)$. We may even go one step further, as we will do here, and consider a *zero-dimensional* field theory. 'That is not much of a theory', you might complain, and your point is well taken. Zero-dimensional means there is no space and no time, so the 'field' has just a single constant mode (like the zero-energy mode of the theory above), so the 'field' is just a real or complex variable. It is very much a toy model that we only introduce to illustrate at a very basic level what the effect of quantum corrections in a field theory looks like.

Our toy model only has two real modes: a light mode with 'mass' m denoted by φ, and a heavy mode with 'mass' $M >> m$ denoted by χ and is defined by a simple polynomial action:

$$S(\varphi, \chi) = \frac{m^2}{2}\varphi^2 + \frac{M^2}{2}\chi^2 + \frac{\lambda}{4}\varphi^2\chi^2\,. \qquad \text{(III.4.11)}$$

This action has no derivatives; the terms quadratic in the fields represent the free modes and the quartic term describes the interaction between the two modes. This very rudimentary theory will in the next section be used to illustrate certain structural (diagrammatic) aspects of perturbation theory and Feynman rules.

- *Quantum Electrodynamics* The third example is the Lagrangian for QED, the theory we discussed already in Chapter I.4 and in the section on gauge invariance in Chapter II.6 ,

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \bar\psi(i\not\partial + m\mathbf{1} + e\not A)\psi\,. \qquad \text{(III.4.12)}$$

Let us make some observations about this Lagrangian:
(i) It is a compact expression of which each part is manifestly Lorentz and gauge invariant.
(ii) Besides the Maxwell field describing the photon, and the Dirac field describing the electrons and positrons, the action contains two parameters: the electron/positron mass m and the coupling constant corresponding to the electron charge e.
(iii) The first three terms are quadratic in the fields and represent the free part of the action. The first term gives rise to the free photon propagator, while the second and third correspond to the free electron/positron propagator. In the Feynman diagrammatic language these propagators correspond to the wiggly and straight lines that were shown in Figure I.4.28 in Chapter I.4, while the final interaction term corresponds to the interaction vertex diagram of Figure I.4.29. They are also shown in Figure III.4.19.

The naive scaling dimensions of fields. To be able to discuss the scaling properties we first determine the naive

Field	Dimension	**Coupling**	Dimension
$\phi(x)$	-1	m	-1
		λ_n	$4-n$
$A_\mu(x)$	-1	e	0
$\Psi(x)$	$-\frac{3}{2}$		

Table III.4.2: *Scaling dimensions in 4-dimensional space-time.* We have listed the naive 'power counting' scaling dimensions in units of length of some fields and coupling constants. The self-couplings λ_n refer to terms in the energy density of the type $\lambda_n\phi^n$. Note that the quartic coupling for scalar field is dimensionless: $[\lambda_4] = \ell^0$.

scaling dimensions of the fields, which are obtained by applying dimensional analysis. A good starting point is the action which in the system with $\hbar = c = 1$ has dimension zero: $[S] \sim \ell^0$, and the Lagrangian therefore has units $[\mathcal{L}] \sim \ell^{-4}$. From the quadratic terms in the Lagrangian of the scalar field given by equation (III.4.16) we learn that the dimension of the field has to be $[\phi] \sim \ell^{-1}$. Consequently the quartic self-coupling of the $\lambda\phi^4$ term λ has to be dimensionless. For the Maxwell field the Lagrangian $[\mathcal{L}] \sim F^2 \sim (\partial A)^2$ and as $[\mathcal{L}] \sim \ell^{-4}$ we conclude that the gauge potential, like the scalar field, scales like $[A] \sim \ell^{-1}$. From the mass term for the Dirac field $\sim m\overline{\psi}\psi$ we obtain that $[\psi] \sim \ell^{-3/2}$. And from the interaction term $e\not{A}\overline{\psi}\psi$ we subsequently verify that the coupling constant e is dimensionless. We summarize the naive scaling dimensions in units of length, of the various fields and coupling constants in the Table III.4.2.

Scaling in classical field theory. Assigning these scaling dimensions to the fields allows us to discuss the scale invariance of classical field theories. To find out we make a scale transformation of the coordinates $x \to \lambda x$. The fields being space-time dependent will transform accordingly, like $\varphi(x) \to \varphi(\lambda x) = \lambda^d\varphi(x)$. Note that the parameters *do not* transform under this coordinate transformation. After the transformation of the coordinates and fields we see that most terms in the action are invariant, and only the mass terms change in the sense that $m \to m' = \lambda m$. The net effect is that after the transformation you get the same theory back but with a different mass parameter. This argument shows that already at the classical level rescaling corresponds to the theory moving through parameter space. A further message is that in classical theories the mass terms break scale invariance. Massless theories like the Maxwell theory are therefore scale invariant. In fact these results also hold for the classical approximation of the quantum theory, where we think of the field excitations as particle states but where we ignore the typical quantum corrections as will become clear shortly. In the quantum domain we have to take into account the inverse relation between length scales (wavelength) and momentum or energy scales. This implies that if we scale the theory by large λ we effectively take the low energy, long wavelength limit which means that the mass is relatively large and in the limit would become the dominant term. In that regime we cannot excite particle modes and there is no dynamics left. If we take the opposite $\lambda \to 0$ limit, then we study the theory in the high-energy regime where the mass effectively plays no role! And at this level of the discussion we would be tempted to conclude that theories become scale invariant in the high-energy limit. However, this conclusion turns out to be premature because taking the quantum corrections into account we will see that these break this naively expected scale invariance.

Quantum complications. To make sensible predictions in quantum field theory that can be compared with experiment, the calculations which are perturbative in nature, require a *renormalization program* to be executed.[4] It is

[4] I must admit that this sounds like the theory is 'abnormal' and has

exactly this renormalization program, which involves cutting off certain momentum integrals that is responsible for the scaling violations. These violations lead to *anomalous scaling dimensions* for the parameters and fields.

We will try to elucidate some of the outstanding features of that program. One we have mentioned already is that rescaling the theory is the same as effectively rescaling the parameters in the model. What one finds is that potentially at every successive step in the quantum approximation new interaction terms may appear in what is called the *effective action*. In other words, conceivable terms that had zero coefficient in the classical theory one starts with, may become non-zero. And the behavior of the theory under rescaling depends on to what extent these extra terms are relevant at the scale one is interested in. The strong requirement of *renormalizability* means that only a finite number of scale dependent renormalizations of parameters and fields is needed to render the calculations finite to any order. This implies that systematic quantum calculations can be made which lead to unambiguous predictions for physical observables to arbitrary precision.

The Euclidean path integral

As we pointed out in the subsection on statistical mechanics in Chapter I.1, there is a interesting analogy between the statistical description of multi-particle classical physics and quantum physics, in spite of all their fundamental differences. This is not too surprising because after all, a field has an infinite number of modes that represents an infinite number of local degrees of freedom, and we learned that quantum field theory defines a Hilbert space with states that can have any number of particles in it.

to go to a camp to be 'renormalized,' through a process of 'ideological purification,' to adapt it to the 'new normal'. This terminology of course started with *normalizing* wave functions and distributions, just meaning imposing a norm, saying nothing about wavefunctions being normal or not.

In classical statistical physics we can derive the thermodynamic properties from the partition function, Z which is the sum or integral over the phase space Γ of the system, weighted by the Boltzmann factor,

$$Z = \int \exp(-H/kT)\, d\Gamma\,,$$

where $H = H(\Gamma)$ is the Hamiltonian of the system, the integral of the energy density over all of space. An important quantity is then the (Helmholtz) free energy F defined as $F = -kT \ln Z$. What we showed in Chapter I.1 was that the *free energy* was equal $F = U - TS$. And we worked through the example of the ideal gas in the section on page 53. One thing is obvious, the (classical) statistical physics underlying thermodynamics becomes racing a dead horse if the temperature is zero, because there is no thermodynamics as everything is stuck in its lowest state. But that is different in the quantum domain.

The analogy. Quantum field theory is basically a theory at zero temperature, though of course a temperature can be introduced in addition. But what makes quantum field theory at zero temperature already interesting is that there are always quantum fluctuations present in the system. This is an unavoidable consequence of the uncertainty principle. Indeed, the role of thermal fluctuations is taken up by the quantum fluctuations, and instead of the temperature the external parameters are typically Planck's constant and possibly some coupling constants. In some sense you could argue that Planck's constant takes the place of Boltzmann's constant and the external parameter that plays the role of temperature is a fundamental coupling strength appearing in the theory. And indeed, whereas the free energy governs the classical phase diagram depending on the thermodynamic variables like P, V, T and S, that role is now played by the masses and coupling constants. Therefore one may expect different quantum phases and phase transitions to occur in different regions of parameter space even at zero temperature.

Statistical Physics		**Quantum Field Theory**	
Phase space	Γ	Field	ϕ
Energy function	$H(\Gamma)$	Euclidean Action	$S[\phi]$
Partition function	$Z = \int e^{-H(\Gamma)/kT}\, d\Gamma$	Path integral	$Z = \int e^{-S[\phi]/\hbar}[d\phi]$
Free energy	$F = -kT \ln Z$	Effective action	$S_{\rm eff} = -\hbar \ln Z$

Table III.4.3: *Correspondence between the fundamental concepts of (classical) statistical physics and quantum field theory.*

The path integral. This fascinating analogy between classical statistical physics and quantum field theory becomes much more tangible once we introduce the *Euclidean path integral* as a tool to do calculations in quantum field theory. In quantum theory we define the (Euclidean) path integral or *quantum partition function*, as a weighted sum over the *classical* configuration space, where each configuration is weighted by the exponential of its classical action:

$$Z \equiv \int e^{-S[\phi]/\hbar}[d\phi]\,. \qquad \text{(III.4.13)}$$

So indeed, the path integral approach to quantum theory does away with wave functions and in fact with Hilbert space, but shows that the same information on quantum amplitudes can be extracted from the corresponding classical expressions, and averaged over all paths or classical field configurations that match the required boundary conditions. Of course, this integration over infinite-dimensional spaces is not simple and to properly define it one encounters a lot of mathematical pittfalls. It requires defining a proper integration measure $[d\Phi]$ for the 'space of field configurations.' But even having a suitable measure, calculating the integral exactly, is too much to hope for, and the best one has been able to do in general is to develop a systematic approximation scheme by expanding the expressions in a power series in $\hbar$ and the coupling constants, using Feynman diagrams and rules. These calculations are notoriously subtle and require a rather unusual arsenal of skills. I will avoid all these highly relevant technicalities here, but nevertheless continue the overall narrative, plainly quoting the results along the way if we need them. And this way I hope to be able to convey the central ideas and discuss what they mean. I refer interested readers to the final section of this chapter where we go a step further in explaining the perturbative approach and consider some specific quantum processes in more detail.

In the comparison with statistical physics the temperature parameter is replaced by some coupling constant times $\hbar$, and S is now the classical(!) (Euclidean) action which is equal to the Lagrangian density integrated over all of Euclidean space-time. The integral involves 'imaginary time', which means that we set $t \to i\tau$, so that the flat Minkowski

space-time just becomes 4-dimensional Euclidean space with $x_4 = \tau$. The idea is that in Euclidean space the mathematical manipulations are much simpler and in particular more convergent than in Minkowski space. But the price one has to pay is that after the calculation is finished one has to 'rotate' back to Minkowski space-time in order to interpret the results.

The effective action. The analogue of a free energy is then the so-called *effective action* $S_{\rm eff}$. :

$$S_{\rm eff} \equiv \hbar \ln Z \qquad \text{(III.4.14)}$$

And as in the definition of Z we have summed or integrated over all field variables, the effective action only depends on the parameters of the theory. This function or its derivatives could become discontinuous, signaling what we have earlier called quantum phase transitions. A strong way to express this analogy is to say that quantum theory in d spatial dimensions is just statistical mechanics in $(d+1)$ dimensions, where the Euclidean action of the d-dimensional space becomes a 'would be' $(d+1)$-dimensional Hamiltonian. An example of this was provided by the $d = 2$ Ising model (discussed in the section on magnetic order in Chapter III.2), where one encounters a quantum phase transition at zero temperature at some critical value of the external magnetic field. It has been shown that the characteristics of that transition indeed correspond to the $d = 3$ classical Ising model.

In Table III.4.3 we have summarized the correspondences between statistical physics and quantum field theory. And it should be said that this Feynman *path integral* approach to quantum theory is in many ways complementary to the operator, Hilbert space approach, and has led to many new and valuable insights into the quantum world. It has become an indispensible tool in our modeling and understanding of physical reality.

Scaling and renormalization

In this section we discuss scaling properties in a generic way, following the renormalization group approach of Kenneth Wilson using the language of the Euclidean path integral and the effective action as introduced in the previous section. We apply the formalism to the ϕ−fourth model. Wilson received the Physics Nobel prize for his work in 1982.

The Wilson approach to renormalization.
The starting point is to define the theory with *momentum cut-off* Λ:

$$Z = \int [D\phi]_\Lambda \exp\left(-\int \mathcal{L}_0 \, d^4x\right), \qquad \text{(III.4.15)}$$

with the ϕ−fourth bare Lagrangian density:

$$\mathcal{L}(\phi, \partial_\mu\phi) = \frac{1}{2}(\partial_\mu\phi)^2 + \frac{1}{2}m^2\phi^2 + \frac{\lambda_4}{4!}\phi^4 . \qquad \text{(III.4.16)}$$

The integration is over all space-time field configurations and has a measure with some momentum cut-off:

$$[D\phi]_\Lambda = \prod_{|k|<\Lambda} d\phi(k) . \qquad \text{(III.4.17)}$$

You can think of it in the following way. Any field configuration can be expended in a complete set of energy-momentum eigenfunctions,

$$\phi(x) = \sum_k a_k \, \phi_k(x).$$

Integrating over the field configurations basically means that you integrate over the space of expansion coefficients, so the measure is then simply:

$$[D\phi]_\Lambda = \prod_{|k|<\Lambda} da_k ,$$

where the integral is only performed over the a_k with $k < \Lambda$. The importance of the cut-off is that all integrals are

calculable in principle but the results may depend on the cut-off.

Integrating out high momentum modes. We continue by splitting the field modes depending on their momentum by defining:

$$\phi = \phi^< + \phi^> \ \text{ with } \ \phi^> = \begin{cases} \phi(k) \text{ for } b\Lambda < k < \Lambda \\ 0 \ \text{ otherwise.} \end{cases} \tag{III.4.18}$$

Now we have to expand out the Lagrangian and split it in the part that depends only on $\phi^<$ that has the same form as $\mathcal{L}$ and the part that depends on both $\phi^<$ and $\phi^>$ and their derivatives. The path integral then becomes a product of two factors. The idea is to perform the integral over the $\phi^>(k)$ components in the second factor.

The effective Lagrangian. The theory obtained after integration over these high momentum modes is an effective theory for the field ϕ but now with a cut-off $b\Lambda$:

$$Z = \int [D\phi]_{b\Lambda} \exp\big(-\int \mathcal{L}_{\text{eff}}\, d^4x\big)\,,$$

where the effective Lagrangian $\mathcal{L}_{\text{eff}}$ will be equal to $\mathcal{L}$ plus an infinite number of correction terms in increasing powers of the coupling constant λ_4 and the field ϕ and its derivatives. The calculation of this expansion is a complicated matter and will not concern us here because the qualitative features we want to address can be discussed without. The philosophy is like a calculation we will do in the toy model in the final section, in that by integrating out a high-mass variable χ we obtain an effective Lagrangian which can be thought of as an infinite power series in the remaining low-mass variable ϕ. In the toy model this can be done explicitly, and therefore gives you a good idea. In the situation here we deal with fields and their derivatives, that all depend on space-time coordinates. The expansion becomes similar to the toy model diagrammatically, but the loop diagrams now involve integrations over the loop momenta in the high momentum range.

Why am I telling you all this, where are we? So far we have mapped a rather simple field theory with a cut-off Λ on a much more complicated theory with cut-off $b\Lambda$. What is that good for? To see that we return to the scaling properties of the terms in the effective Lagrangian, and apply dimensional analysis to the new interaction parameters introduced by integrating out the high momentum modes. Let us write,

$$\mathcal{L}_{\text{eff}} = \mathcal{L}_0 + \text{correction terms}\,,$$

where $\mathcal{L}_0$ only contains the quadratic terms describing the original free field theory. The correction terms in principle contain all powers of the field ϕ and their derivatives. This is somewhat disturbing as we now have to deal with an extremely complicated effective description, but we are not done yet.

The effective theory has a momentum cut-off $b\Lambda$, and it is this theory we want to rescale. We do so by rescaling the momentum by $k \to k/b$ and the coordinates by $x \to bx$. This rescaling of the coordinates brings out certain powers of b in front of the terms in the effective Lagrangian. And because $b < 1$ we are going to smaller spatial and larger momentum scales. As the Lagrangian has dimension ℓ^{-4} and $[\phi] \sim \ell^{-1}$, one finds that the coupling constants $g_{m,n}$ for a term with a power of the field m and the number of derivatives equal n has to scale with a power of the scaling factor b given by:

$$g_{n,m} \to b^{n(d/2-1)+m-d} g_{m,n}\,.$$

This expression is consistent with the values we assigned before, for example the ϕ^4 coupling $\lambda_4 = g_{4,0}$ in a space-time dimension $d = 4$ yields indeed a power equal zero, confirming that λ_4 is dimensionless. Now we want to distinguish three possible cases for the scale dependence of the couplings in the effective Lagrangian:

– power > 0: the term is *irrelevant*
– power $= 0$: the term is *marginal*
– power < 0: the term is *relevant*.

At this point we should mention that also extra terms of the type that were already present in $\mathcal{L}_0$ will be generated and these terms are absorbed into the new renormalized fields and parameters. In the case at hand $\phi \to \phi'$, $\mathrm{m} \to \mathrm{m}'$ and $\lambda_4 \to \lambda_4'$, so that indeed $\mathcal{L}_0$ in the next iteration looks the same but with renormalized fields and parameters. If we imagine iterating this procedure, repeating the rescaling after integrating out the highest momentum modes, we get a sequence of maps of the coupling constants, and it is this sequence of maps that we refer to as the *renormalization group trajectory*. What we arrive at is a flow of the model in the space coupling constants. The important point of distinguishing the various terms is that the irrelevant ones get suppressed by powers of b, while the relevant ones have inverse powers and will grow. For the marginal operators one has to make higher order calculations to determine which way they will go. The upshot is that the renormalization group action maps out a trajectory of the given theory in the space of coupling constants, in other words, in the space of theories. In the next section we will discuss the *renormalization group equations* that determine the trajectories and go through some relevant examples.

Note that the question whether interaction terms are relevant, irrelevant or marginal depends strongly on the space-time dimension d. One can easily check that the ϕ^4 term is marginal for $\mathrm{d} = 4$ but it becomes relevant if $\mathrm{d} < 4$. For $\mathrm{d} = 2$ one finds that all powers of the field become relevant, because the exponent becomes -2 for all of them. A mass term scales as expected like $\mathrm{b}^{2-\mathrm{d}}$ and is therefore relevant for all $\mathrm{d} > 2$.

The asymptotic behavior of the theory one considers now depends on where these trajectories go. They may move towards a fixed point that could be either zero or nonzero, or trajectories could run off to infinity, which means that the theory looses its meaning and becomes inadequate to describe the physics. The irrelevant terms go to zero as they are suppressed by the increasing powers of the cutoff. So most of the scary looking terms that appeared after integrating over high momentum modes disappear again because of the rescalings, and because of their irrelevance. This brings us back to the question of scale invariance. If the couplings in a theory go to a fixed point, then the theory defined by that fixed point is by definition scale invariant!

We note that the ϕ^4 theory in four space-time dimensions has what is called 'trivial' fixed point where the parameters m^2 and λ_4 are both zero, and $\mathcal{L}^{(\prime)} = (\partial_\mu \phi)^2$. This theory is in fact invariant under the conformal group as we have mentioned before. It has been shown that the ϕ^4 theory for $\mathrm{d} = 3$ has a non-trivial fixed point, The so-called Fisher–Wilson fixed point.

The statement is that theories that have only relevant and marginal terms are called *renormalizable*. It is in those theories that it is possible to take the cut-off Λ to infinity sending the irrelevant terms to zero. The effect of all the quantum perturbations can then be absorbed in sensible redefinitions of field and parameters.

The importance of the Wilson's renormalization group perspective is that it a priori assumes that there is a real physical cut-off and that the physics at lower energy may show some dependence on it. This typically is the case in applications in condensed matter and you had better take it into account. There is no need to send the cut-off to infinity, because it is really there. On the other hand it used to be somewhat of a mystery if not a miracle why the fundamental theories like the Standard Model are all renormalizable (from the start). And one wondered why Nature was so judicious in its choice. Just to please physicists so they could do meaningful perturbative calculations? The Wilson approach makes clear that renormalizability is exactly what survives in a natural way. Those are the terms that basically survive in the renormalisation group flow. Quite arbitrary theories may well flow towards a scale invariant fixed point that lies inside a subspace of relevant renormal-

izable theories, which do not need to be scale invariant! The Wilsonian perspective we have outlined leads to the conclusion that the renormalizable models are universal in that they describe the asymptotic behavior of large classes of other models. ■ ■

The quantum bank. Whether you study the stars, write poems, or are world champion armwrestler, in the end we all have to deal with banks. You need a loan or a mortgage and you get immersed in a labyrinth of options: this one looks even more advantageous then the other. Ultimately it always boils down to interest rates, and those rates are calculated based on a mysterious mixture of facts and fictions concerning the certainties of your present and the uncertainties of your future. But one thing remains true under all circumstances: borrowing money costs money! And you are happy because you are spending money you don't have!

Now back to quantum, In the realms of quantum theory the currency is energy rather than dollars. Yet there is also a bank, which is basically the vacuum itself. We know that because of the Heisenberg uncertainty relations, a quantum marble cannot be at rest at the bottom of the bowl, it has to jiggle around a bit. There is no certainty ever in the quantum world. This may work to the advantage of the participants in the sense that there are always quantum fluctuations even in the ground state and even at zero temperature. Quantum reality is such that there is always some energy around. And the idea of the cooperative quantum bank is that it provides very cheap energy loans, but they come with some unusual restrictions. The slogan is, you can borrow as much as you want but only for a very short period.

Whereas the money banks usually have very high interest rates for ultra short-term loans, the quantum bank's energy loans work exactly the opposite way. As long as you $\Delta E x \Delta t \leq \frac{\hbar}{2}$ you are doing fine. So if I am a photon and play it big, I can borrow energy so that I can produce for example an electron-positron pair to impress my fellow photons as long as the loan is very short term. But now the catch is that because the overall energy has to be conserved, the quantum bank insists that you return your energy before the Federal Reserve gets wind of it. And this is what certain real-life Quants in real banks don't seem to understand. There is a moment of reckoning: you speculate yourself into heaven, but you have to be back home with two legs on the ground in time! In other words the quantum world makes sure that the pair just created annihilates back into the vacuum and the photon continues its journey, as if nothing ever happened to it. You would think. But no it isn't as simple as that. The photon carries its creative banking experiences with it and they effectively change its behavior.

It reminds me of my good old student days at Delft University, when I was cycling home late at night along the beautiful 'Oude Delft' canal from the lab, or was it a party? Suddenly I got pulled over by the police. Trouble! Probably a costly ticket because I had no lights on my bike. And while the officer was searching for his ticket book in the car, I shoved my old bike in the canal. Bloop...gone! When the police officer returned and started to make a solemn declaration about 'your bike sir appears to be missing some appropriate lighting'....I interrupted him and asked what bike he was talking about. 'But I thought that ...' 'Yes, may be you thought, but look ...' This caused some consternation. Indeed, here were powerful fluctuations at work that the officer on duty apparently had no working knowledge of! □

Running coupling constants

As we have seen, quantum theory and in particular quantum field theory has come up with a surprising answer to questions about the spatial or momentum scale dependence of the coupling parameters in a given theory. Though the road to the result is highly technical and the arguments may at first appear to be quite opaque, what results is clearcut and strikingly simple.

The renormalization program yields equations that govern the behavior of the parameters of the theory as a function of scale. These are differential equations that remind you of an ordinary dynamical system, say a set of interacting Newtonian particles. Now you have to imagine that these equations describe the 'motion' of a given theory in parameter space, not as a function of time but as a function of scale! And that's where the term 'running coupling constants' comes from. It is kind of mind boggling to think of a given theory 'running' in the space of theories. Yet that is what happens and moreover, it teaches us about the limiting or asymptotic behavior of such theories. This may be in the high momentum (ultraviolet) or the low momentum (infrared) limit, depending on the problem one is interested in.

The first and maybe simplest equation of this type – called the Gellman-Low equation – was written down for QED. Later the general renormalization group approach which we described, culminated in the so-called Callan-Symanzik equations for the scaling behavior of any composite local operators of the type we encountered in the expansions. These *renormalization group equations* govern the flow of points in the space of (renormalized) coupling constants which we will denote by $\mathcal{W}$. Let us consider the simple case of a single coupling constant g. The theory has a momentum cut-off Λ, and the equation involves the renormalized coupling which we denote by $\bar{g}$, which depends on the momentum scale through its logarithm only, $\bar{g} = \bar{g}(\log \bar{p})$, where we choose $\bar{p}$ to be the dimensionless momentum variable $\bar{p} \equiv p/\Lambda$. The renormalization group equation has the simple form:

$$\frac{d\bar{g}}{d\log \bar{p}} = \beta(\bar{g})\,. \tag{III.4.19}$$

This equation just says that the rate of change of the coupling $\bar{g}$ equals a function $\beta(\bar{g})$, not surprisingly called the *beta-function*. This function depends on $\log \bar{p}$, but only through the coupling constant $\bar{g}$. In that sense you can think of it as a *functional equation* for $\bar{g}$ as a function $\log \bar{p}$, in the spirit of equation (III.4.5).

Let us assume that at some large distance (small momentum) this coupling is small, then we may look at β for small $\bar{g}$ and develop it there as a power series like:

$$\beta(\bar{g}) = a\bar{g} + b\bar{g}^2 + \cdots\,, \tag{III.4.20}$$

and for small $\bar{g}$ the successive terms will become ever smaller, and we can safely truncate the series. Now given the quantum field theory the coefficients $a, b, \cdots$ can be calculated using perturbation theory. This approach allows us to deduce important general features of the theory. It is important though to note that because the beta function is mostly calculated perturbatively, it follows that the results obtained can only be trusted in the domain where the perturbation theory holds, in other words, where the expansion parameters are small. Of course, the rare cases where models can be solved exactly serve as ideal testing grounds for the tools we are describing here.

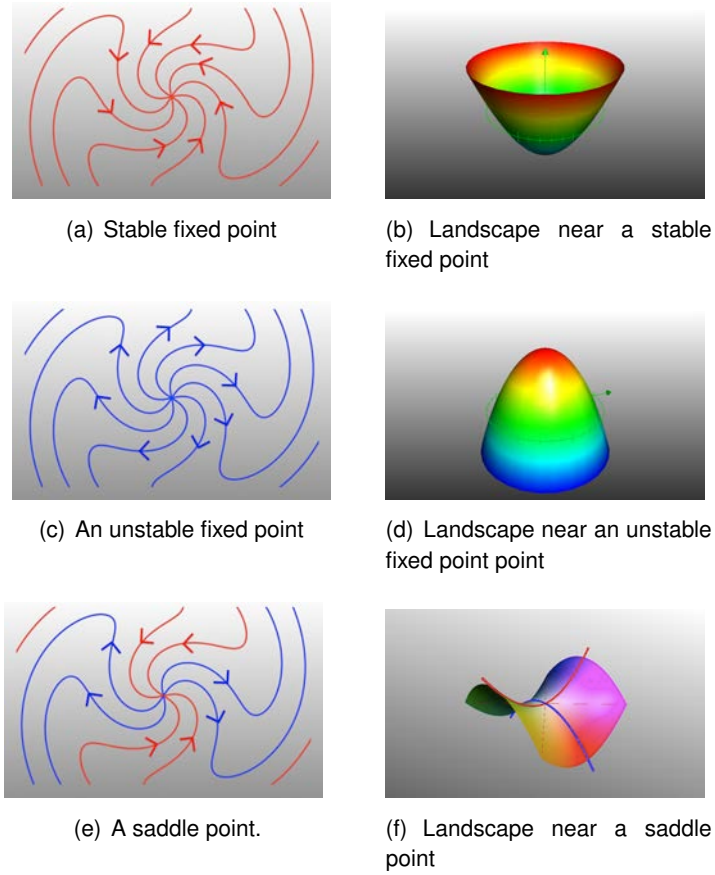
Mechanical analogues

Let me point out a mechanical analogy that should be familiar and thus helpful. It refers to our discussion on dynamical systems on page 11 of the section on Newtonian mechanics in Chapter I.1. If we think of a complicated theory with many parameters, we will have a system of coupled equations but all with the same first derivative with

respect to $\log \bar{\mathrm{p}}$ on the left-hand side. Using $\log \bar{\mathrm{p}}$ instead of $\bar{\mathrm{p}}$ makes the equations particularly simple, the only thing you have to keep in mind is that $\log \bar{\mathrm{p}}$ grows monotonically with $\bar{\mathrm{p}}$, so if $\log \bar{\mathrm{p}}$ becomes large then $\bar{\mathrm{p}}$ does also, but for $\bar{\mathrm{p}}$ going to zero $\log \bar{\mathrm{p}}$ goes to minus infinity. In this sense we may think of $\log \bar{\mathrm{p}}$ as some kind of 'time' variable t. Then the equation just describes the motion of a point in the (coupling constant) space $\mathcal{W}$ of the system.

Stated differently, as the point represents a particular theory, its motion describes a trajectory in a space of theories! The left-hand side is the 'velocity' or rate of change which depends – through the expression on the right-hand side – on where you are in the coupling constant space. So, the equation defines a vector or flow field over $\mathcal{W}$, in a similar way that Newton's dynamical equations for a particle define a flow over the phase space, as we discussed in Chapter I.1. The equation governs the trajectories completely once the initial conditions for $\bar{\mathrm{g}}(\mathrm{t})$ at some $\mathrm{t} = \mathrm{t}_0$ are given, just like Newton's equations do after you give the initial positions and velocities of a bunch of interacting point particles. These dynamical systems are usually nonlinear, and also in the case at hand the dynamical system is nonlinear as we see from the expansion in equation (III.4.20). As we remarked in Chapter I.1, it allows us to search for universal behavior, because the system may for large values of time, or high momentum ($\mathrm{t} = \ln \bar{\mathrm{p}}$), end up in some fixed point or limit cycle and may for long time scales exhibit universal behavior.

It is amusing to see how we manage to address deep questions in the realm of Quantum Field Theory because we have been able to map the problem onto a rather simple Newtonian dynamical system. Indeed, from equation (III.4.19) one sees that for the points where the β-function vanishes the 'velocity' is zero, so these points correspond to stationary points. This translates into the statement that that theory becomes invariant under further rescaling. It is a theory which is called ultraviolet (high momentum) stable, because it has ended up in in some scale invariant fixed point. Note that this analysis allows for theories that are quite different initially to end up in the same ultra violet fixed point. They belong to the same universality class.



(a) Stable fixed point

(b) Landscape near a stable fixed point

(c) An unstable fixed point

(d) Landscape near an unstable fixed point point

(e) A saddle point.

(f) Landscape near a saddle point

Figure III.4.15: *Fixed points.* The three types of fixed points one may encounter in a two-dimensional parameter space. In (f) we see that lines of steepest descent and steepest ascent are perpendicular.

For a single coupling we have only one dimension and it is straight forward to see what is possible for small coupling, where we only take along a few terms in the expansion (III.4.20). For example, the system may move to a stable fixed point where it would stay for ever after, or we may have an unstable fixed point and the system would

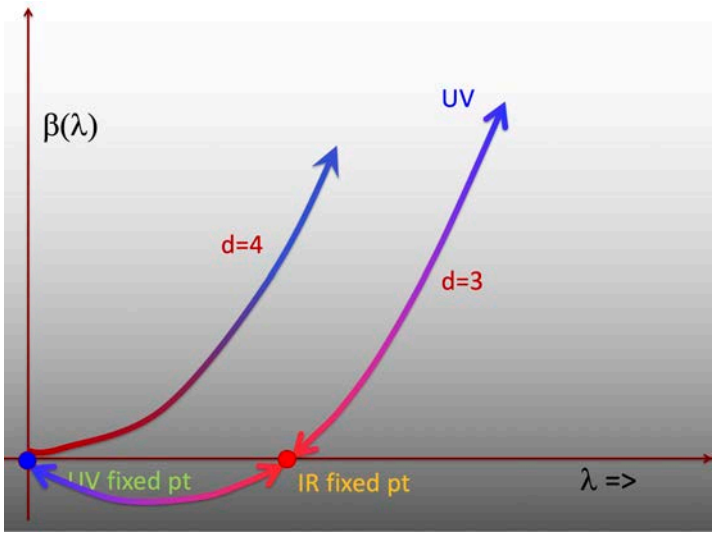


Figure III.4.16: *Beta functions of the ϕ^4 theory in 3 and 4 dimensions.* Depending on the starting value of λ_i we have sketched the asymptotic behavior of the coupling constant $\lambda = \lambda(\log p)$. In the infrared region (decreasing $\log p$) there is a non-trivial IR fixed point for $\lambda = \lambda_0$ for $d < 4$. In the ultraviolet limit (increasing $\log p$) we find for $\lambda_i < \lambda_0$ a trivial fixed point ($\lambda = 0$) where the theory behaves like a free theory with zero coupling. For $\lambda_i > \lambda_0$ the coupling keeps increasing, at least for as long as the expansion makes sense.

move away from it under any small perturbation.

If we think of a two-dimensional parameter space, such stationary points can in general only have three generic types of behavior: either the point is attractive, or repulsive, or it is a saddle point. This is illustrated in Figure III.4.15. In two or more dimensions one also could imagine the possibilities of limit cycles, or more exotic attractors where the system could display even chaotic behavior. But generic features of these renormalization group equations appear to exclude that. Nature saves us the humbling demise that our theories would get lost in chaotic asymptotics. Coping with quantum uncertainties is enough of a challenge!

The scalar ϕ^4 theory. Let us now turn to an explicit example. What do the renormalization group equations look like for the scalar model we have been discussing? It has two parameters, $\bar{m}^2$ and $\bar{\lambda}_4$, and therefore two equations with two beta functions. To lowest non-trivial order these read as follows:

$$\frac{d\,\bar{\lambda}_4}{d\log\bar{p}} = \beta_\lambda(\bar{\lambda}_4) = -(4-d)\bar{\lambda}_4 + \frac{3\bar{\lambda}_4^2}{16\pi^2}\,, \qquad \text{(III.4.21)}$$

$$\frac{d\bar{m}^2}{d\log\bar{p}} = \beta_m(\bar{\lambda}_4) = [-2+\gamma_m(\bar{\lambda}_4)]\bar{m}^2\,. \qquad \text{(III.4.22)}$$

Let us make some observations with respect to these equations:
(i) The constant term on the right-hand side of the equations gives the naive scaling dimensions, namely 0 and -2 respectively.
(ii) The other terms are radiative corrections to the numbers, and are supposed to be small. The anomalous term $\gamma_m(\bar{\lambda}_4)$ vanishes if $\bar{\lambda}_4 = 0$.
(iii) As all corrections take the form of a power series in $\bar{\lambda}_4$ only, it is the $\bar{\lambda}_4$ equation that drives the dynamics. So, let us then start with the first equation. In four or more dimensions the beta function is positive and the coupling will therefore keep growing until it becomes so large that the perturbation series breaks down in that successive terms are no longer decreasing. What happens in the strong coupling regime in that case cannot be answered through this analysis, because the series diverges the approximation scheme becomes invalid. One would have to resort to strong coupling approximations meaning numerical lattice simulations. The conclusion appears to be that there is no fixed point at larger values of the coupling, which means that the theory deteriorates into the quartic term, not a physically interesting or meaningful result.
(iv) In Figure III.4.16 we depicted the beta function $\beta_\lambda(\bar{\lambda}_4)$ for $d = 3$ and $d = 4$. Where the blue direction is the direction of increasing momentum (ultraviolet), while the red arrows point in the decreasing momentum (infrared). We see that for $d < 4$ the beta function has two zeros, meaning that there are two fixed points: one at zero and one

larger than zero at some value λ^*. The new point is an infrared stable fixed point.
(v) If we let d approach 4 from below we see that the two fixed points merge at the trivial fixed point $\bar{\lambda}_4 = 0$ which corresponds to the free field theory.
(vi) Finally, the renormalization equation for the mass parameter has on the right-hand side the constant -2 which just reflects the naive scaling we have discussed already. If we scale up the theory one expects the mass term to become less and less relevant. while in the trivial fixed point it is the one and only relevant parameter. The precise form of the solution is:

$$\bar{m}^2 \simeq \left(\frac{m}{\Lambda}\right)^2 \left(\frac{\Lambda}{p^2}\right)^{2+\gamma} . \qquad \text{(III.4.23)}$$

If in less than four dimensions the system sits in the nontrivial Fisher-Wilson infrared fixed point, we get the anomalous correction to the naive scaling law corresponding to $\gamma(\lambda^*)$. This correction plays a role in the $d = 3$ statistical physics of magnetic materials. It has no effect on the ultraviolet behavior of the theory.

Gauge couplings

The following picture emerges: the constant a in the expansion of the beta function (III.4.20) has a generic structure:

$$a = d - n , \qquad \text{(III.4.24)}$$

where d is the physical space-time dimension and n is some critical dimension, critical because the scaling behavior of the theory depends critically on whether d is smaller or larger than n. If $d < n$ then $a < 0$ and the growth rate of g is negative and g will decrease with growing momentum, or what amounts to the same, with decreasing distances. In this case the coupling constant will go to zero, its like no interactions are left at small distances. And as the linear approximation will get better and better with decreasing g, the prediction that this theory behaves as a theory of free particles at small distances is reliable and consistent.

If however $d > n$, the situation looks pretty bad because now the coupling grows bigger at smaller distances and the approximation breaks down and we would need the complete β function.

Now there is still the 'in between' possibility with $d = n$, and it turns out to be of considerable interest in the situations that nature faces us with. In that case we have to turn to the next term in the series with coefficient b. If we only keep the b term the solution becomes:

$$g(p) = \frac{c}{1 - bc \log p} , \qquad \text{(III.4.25)}$$

with c some positive constant. Again, we may look at what happens for when b is positive, respectively negative. The different behaviors are plotted in Figure III.4.17 and interestingly we encounter both cases in realistic particle theories.

Quantum electrodynamics. The case with positive b corresponds to pure *quantum electrodynamics (QED)*, the theory of photons, electrons and positrons. From the blue curve we see that g becomes very small for small values of $\log p$, that is large distances. We recall that the expansion parameter in QED is the fine structure constant $\alpha = e^2/4\pi\epsilon_0\hbar c \simeq 1/137$. This corresponds to the familiar regime where charges are free and have weak electric interactions, and perturbation theory can be trusted, and allows for calculations of exceptional precision. The decrease of the coupling at larger distances reflects the situation that the quantum fluctuations tend to screen the 'bare' or 'naked' charge. This effect is called *vacuum polarization* because due to the quantum uncertainties the quantum fluctuations in the fields result in the excitation of virtual electron-positron pairs and these screen the 'bare' charge. Vacuum polarization is discussed in more detail in the following section on page 579.

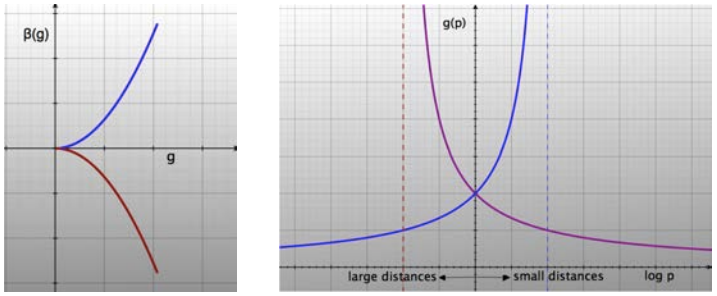


Figure III.4.17: *Running coupling constants.* On the left a plot of the beta function $\beta(g)$ of equation (III.4.20) for positive (blue) and negative (purple) sign of the constant b. On the right we plotted the solutions for $g(\ln p)$ of equation (III.4.25), showing the dependence of the coupling strength on the logarithm of the momentum. The blue curve goes to infinity for a finite value of $\log p$ at the so-called Landau singularity. The purple curve corresponds to negative b, the coupling tends to zero for small distances which is called *asymptotic freedom*.

For increasing momenta the coupling becomes infinite for some finite momentum scale, which taken literally would suggest that the naked charge would be infinite. This singular behavior is called a *Landau pole*, and the presence of such a pole indicates that the theory becomes untenable past a certain scale and has to break down somehow. In general, it is true that a coupling growing large is a strong signal that the theory is no longer to be trusted past that point. This is not a disaster but just a whistleblower announcing that the model is losing its validity and presumably some new physics has to enter the conversation to allow us to escape the singularity.

This illustrates again a notion that I have mentioned before, namely that theories are not right or wrong per se, but rather have a limited domain of validity. In the present context a large coupling usually means that the physical system will enter another regime for which the theoretical picture one started off with becomes inadequate. Renormalization in that sense helps theories to predict their own demise. How nice to have theories which know about their own limitations. For the case at hand the resolution came much later when it was discovered that at small scales it made no sense to look at the electromagnetic interactions separately. The remedy was to combine the electromagnetic and the weak nuclear interactions into a single unified 'electroweak' theory which turned out to behave extremely well also for extremely small distances as we have been able to verify in the *Large Hadron Collider* (LHC) at the European accelerator center CERN in Geneva. In fact, the word 'large' here implies precisely 'large momenta', and this collider smashes particles into each other with very high energies, and that means that they can come very, very close to each other. The LHC was specially built to investigate what happens to the interactions at very small distances.

Quantum chromodynamics. Let us now look at the purple curve in the figure corresponding to negative values of b. It shows that the coupling goes to zero with increasing $\log p$ or at smaller distances. Therefore, the theory ends up describing non-interacting – free – particles for large momenta. This behavior under scaling is realized in *Quantum Chromodynamics (QCQ)*, the theory for the strong nuclear force. We see that the 'strong' interactions between quarks, paradoxically enough becomes extremely weak at small distances. This remarkable behavior of the strong interactions is called *asymptotic freedom*. For a long time it was thought that the problem of the strong nuclear forces could never be solved along the lines of quantum field theory, but this picture changed drastically after 'asymptotic freedom' was discovered and the strong interactions were tamed because they turned out to be the manifestation of a well-behaved weakly coupled theory at small distance scales. This totally different asymptotic behavior of QED and QCD, is of course due to the self-interacting nature of the gluons. Those self-interactions distinguish the non-abelian from the abelian theories. For the discovery of asymptotic freedom the physics Nobel prize 2004 was awarded to David Gross, David Politzer and Frank Wilczek.

From the purple curve we also see that going towards small momenta the coupling grows 'without limit' at some finite value of $\log \mathrm{p}$. This behavior is sometimes called *infrared slavery* because the particle would become extremely strongly coupled. The physical interpretation of an increasing coupling constant is that at a scale where the coupling becomes of order unity, the perturbative predictions lose their reliability, and one expects other physics and non-perturbative effects to come into play. For QCD there are two fundamental phenomena that are linked to this. The first is the formation of the quark-antiquark condensate that causes *chiral symmetry breaking* as we discussed in Chapter II.6 on symmetry breaking on page 441. This symmetry breaking lead to the interpretation of the three pion particles ($\pi^{\pm}$ and π_0) as the 'massless' Goldstone degrees of freedom associated with the breaking. The second non-perturbative phenomenon manifest at that scale is the confinement of quarks. As mentioned before the collective of quarks reorganizes itself into tightly bound composites called *hadrons* made up of either three quarks (called *baryons*) or form a quark an anti-quark pair (called *mesons*). The protons and neutrons are the nuclear particles from which all familiar forms of matter are build, and these are baryons. The pions however belong to the group of the mesons. What this means is that at scales where the becomes large the perturbative approach breaks down and the behavior of the theory is no longer what one would expect from the its weak coupling behavior. At that point it is important to switch to a different effective theory that is formulated directly in terms of the hadrons. In the case of QCD this turns out to be a *nonlinear sigma model* that we will not further dwell on.

Grand unification: where strong joins weak

The idea of renormalization and running coupling constants led to a powerful insight into the possibility of unifying the different types of interactions into a single framework often

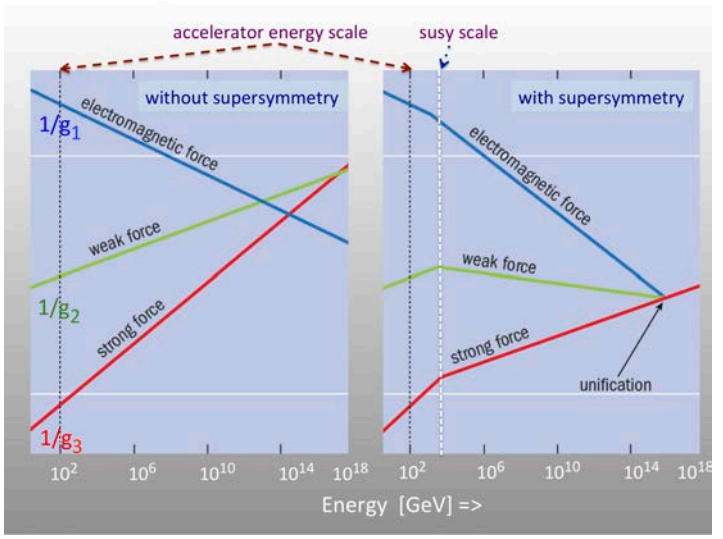


Figure III.4.18: *Unifications.* The subsequent unifications of fundamental interactions suggested by the running coupling strength of the various forces meeting at ever higher energy scales. Experiments at the Large Hadron Collider at CERN in Geneva go up to about 10^3 GeV.

referred to as a *Grand Unified Theory* or GUT. We have already alluded to the fact that the problem of the ill-defined electromagnetic coupling at small distances was resolved by the unification of the electromagnetic and the weak interactions. On the other hand we mentioned the strong nuclear force which turned to become weak at short distances. As we explained in Chapter I.4, these three interactions are now described in a single combined theory called the *Standard Model*. This theory has so far successfully survived extensive testing through many different types of experiments and appears to be able to predict and explain all the data that are available at present.

To give you an impression of what could be a successful next step up the quantum ladder of unification you should look at Figure III.4.18. The picture gives the expectation of how the grand unification could be achieved. Experiments go up to a level 500 GeV so we have witnessed the electroweak unification and we see the strong coupling com-

ing down. Applying the renormalization techniques and scaling arguments we discussed before to the Standard Model, one may calculate the trajectories of the various coupling constants (assuming that no new physics shows up at other intermediate scales) to substantially higher energy scales and indeed it is suggestive to anticipate a further unification at the GUT scale of around $10^{15}\ GeV$. In fact if we extend the Standard Model to its minimal supersymmetric extension, the resulting trajectories for the three couplings of the model really intersect at a single point near $10^{15}\ Gev$ as is shown on the right-hand side of Figure III.4.18. Then the extra (susy) scale of the breaking of supersymmetry has to be introduced because we haven't observed any superpartners of the ordinary particles at low energies. Even more speculative would be the unification with gravity at the Planck scale $10^{19}\ GeV$. Such are the grand vistas and holy grails of modern high-energy physics.

Phase transitions

In the previous sections we have seen that in many body systems described by statistical mechanics or quantum field theory, we may by changing the external parameters being the temperature or some coupling constants have the theory end up in a fixed point of the renormalization group equations. In points where the $beta$-function vanishes the theory is scale invariant. We have seen an ultraviolet fixed point in QCD and an infrared fixed point in the ϕ-fourth theory in three dimensions.

In most of physics this remarkable property of scale invariance is the hallmark of a so-called *critical point* where the system exhibits critical behavior. The behavior around such fixed points, may exhibit fluctuations on all scales, but these can be understood because of the self-similar nature of their spectrum. The correlations display a power law behavior.

The power laws that characterize the critical behavior have universal properties which only depend on the dimensions and nature of the critical point. Many models which may be much more complicated for example having quite a few parameters at the start may move into a universal fixed points where their behavior is described by a much simpler model with fewer parameters. In many lower-dimensional cases the critical models can be solved exactly, which provides important insights about the phase structure of large classes of models, think for example of the Ising model. These ideas where initially developed in statistical physics by Michael Fisher and Leo Kadanoff, and as mentioned in the context of quantum field theory carried further by Kenneth Wilson who received the Nobel prize for his work in 1982.

And it is indeed by the *renormalization group* approach that theorists have on the one hand been able to come up with many interesting and successful explanations, and on the other have been able to construct representative models for a myriad of physical phenomena. They could solve these simplified models exactly and therefore could provide calculable models for a vast range of critical phenomena.

So to conclude, we have shown that one may think of a space of coupling constants where a given theory is characterized by some point in that space where the couplings take particular values. Now there is a set of coupled renormalization group equations for this set which determines a flow of the point through this space that may or or may not end up in some fixed point. In a fixed point the system's behavior becomes scale invariant, and as such it exhibits some characteristic universal behavior of the theory. The renormalization group equations define flow lines in the space of parameters and starting at a given point in the space the theory follows the flow line to some fixed point. Clearly many different theories can end up in the same type of fixed point and that is what we mean by universal critical behavior see Figure III.4.15(a).

On the calculation of quantum corrections

Renormalization is a scheme that guarantees a peaceful coexistence with infinities.

Perturbation theory

What do we mean if we say we have a quantum field theory like QED or the standard model, or a theory of pions, or of superconductivity? The casual term 'The theory' usually refers to a number of inclusions or formal steps starting from three ingredients:
(i) an *action* (or Hamiltonian), which allows us to derive a set of
(ii) *Feynman rules* describing the propagators (two-point (correlation) functions) for the particles in the theory, and also the interaction vertices;
(iii) If we are interested in a particular physical quantum process, we can usually *not* calculate the probability amplitude for that process exactly. It is however possible to make a systematic perturbative approximation, by making a diagrammatic expansion for the quantum amplitude of any process in increasing powers of the relevant coupling constant(s) and in powers of $\hbar$.
Such an approximation scheme is only reliable if the expansion parameter is sufficiently small. This procedure, called *perturbation theory*, is schematically depicted in Figure III.4.19.

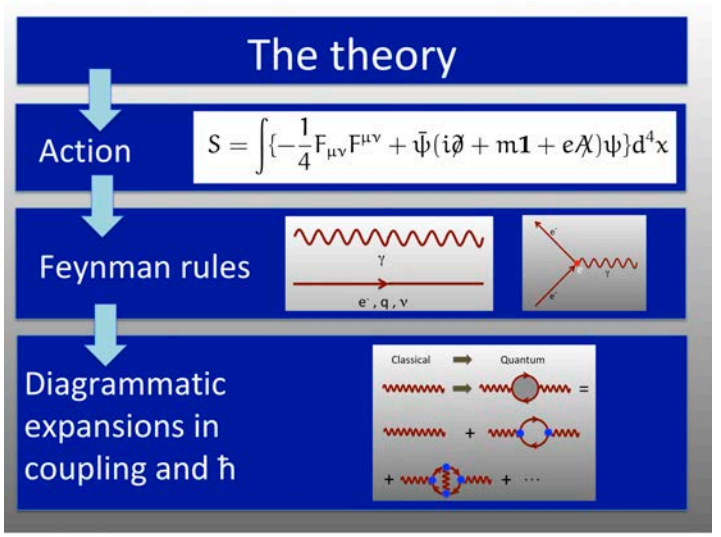


Figure III.4.19: *The perturbative approach.* A theory (like QED) is defined by its classical action, giving the functional form of the theory in terms of the particles (fields) and their interactions. From the action one derives the set of Feynman rules that allow for a systematic diagrammatic expansion of any physical process, This is a series expansion in increasing powers of the coupling(s) and the Planck constant $\hbar$.

The toy model as tutorial in the language of diagrams. Let us take a very simple toy model to illustrate the quant- essential difference between classical and quantum reasoning.[5] The model concerns a drastic simplification and only serves to illustrate certain generic properties of quantum corrections. We are not about to really calculate anything realistic because it turns out that those calculations are quite complicated, and it is where a lot of bright students spend a considerable amount of time on. But luckily we are not the part of the workforce we are just curious tourists! We just want to stare in awe at the statue and need not make one ourselves; we love to eat a sausage but rather not go through how they are made! We are here to see how others did the work!

The toy model is a field theory in zero-dimensional spacetime, where we consider two real valued fields φ and χ. You could say we are studying a system with two modes. The action function has only mass terms and an interaction term (there are no space or time derivatives) and looks therefore almost trivial:

$$S(\varphi,\chi) = \frac{m^2}{2}\varphi^2 + \frac{M^2}{2}\chi^2 + \frac{\lambda}{4}\varphi^2\chi^2\,, \qquad \text{(III.4.26)}$$

and let us assume that $M \gg m$. You might wonder, what if

[5] I encountered this model in a set of lecture notes on 'Applications of QFT to Geometry' by Dr Andy Neitze of Princeton University.

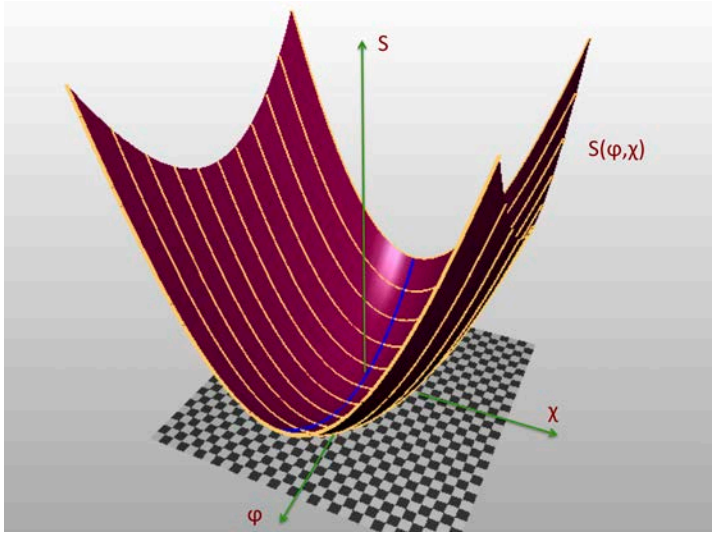


Figure III.4.20: *Action of toy model.* The surface corresponds to the classical action $S(\varphi,\chi)$ as a function of the variables φ and χ.

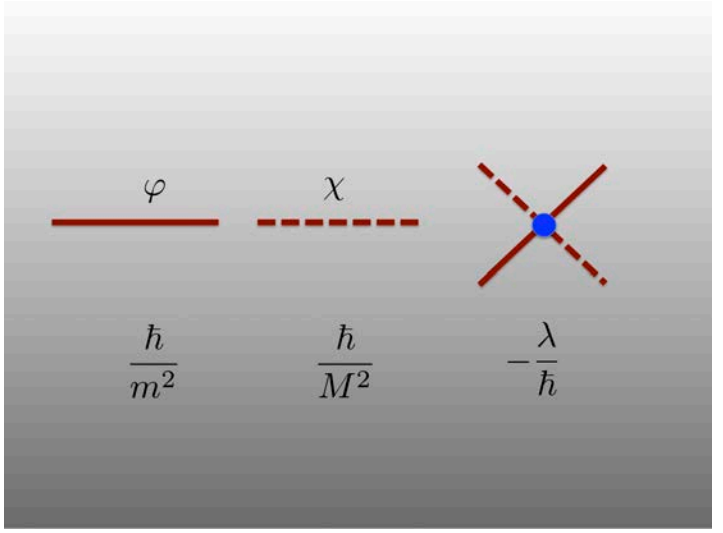


Figure III.4.21: *Feynman rules for toy model.* The Feynman rules are derived from the action and give the functional for the various terms in the action.

anything we can learn from a model in which such a drastic amputation of reality has taken place. In a sense you are correct: the fields are just real variables, and the quantum aspect as we will see is kind of restricted to the $\hbar$ which we stick in. So, in the end we integrate a function, and expand the result, and yes the structure one obtains looks very much like the things we encounter in field theory. This is a pedagogical workout, illuminating and even fun. Let us therefore respectfully execute some 'standard calculations' imagining that we are dealing with a real field theory and see what it delivers and also what not.

The three terms in the action correspond to the three Feynman rules (the elementary diagrams) that we give in Figure III.4.21. The free part yields the two 'propagators,' and the quartic term yields the interaction term with coupling strength equal λ.

Effective actions. Let us consider the most trivial process imaginable namely where the in and out state are both empty. Classically if nothing goes in and nothing goes out then there is a unit probability that nothing happens in between.

What we have is a very heavy mode and a very light mode that interact with each other. What you have classically is that it costs a lot of energy to excite the heavy χ mode and that it is easy to excite the light φ mode. So, for energies well below M only the light mode will be present and we can forget about the heavy χ mode altogether. But if we do a quantum calculation, we should allow for virtual manifestations of the heavy mode, and we have to integrate over all possible values that field may take. We say that we 'integrate out' the heavy mode. And this in turn will drastically change the resulting effective theory for the light mode. It will change three things: (i) it will change the mass of the light mode, (ii) it will change the strength of the interaction term and (iii) it will generate an infinite number of new self-interactions for the light mode. These are quantum effects that affect the low energy behavior of the theory. And these are precisely the generic aspects we like to illustrate with this tiny toy model. We can integrate over the the χ

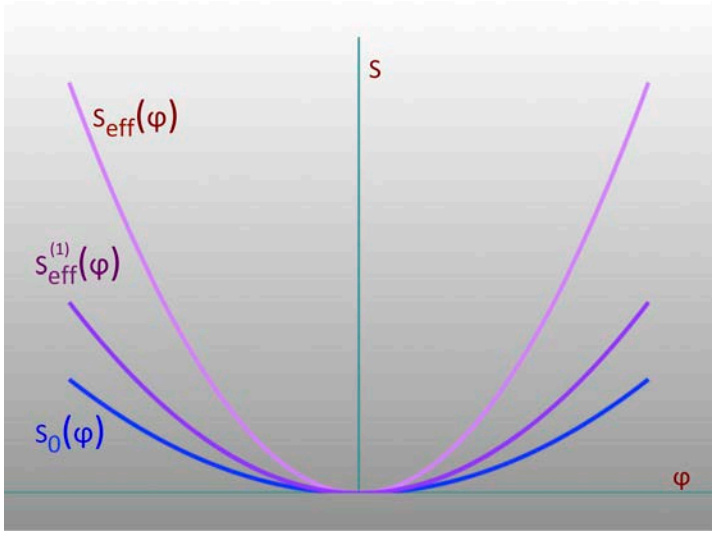


Figure III.4.22: *Effective action for φ field.* The graphs correspond to, (i) the classical action $S_0(\varphi) \equiv (\varphi, 0)$ (blue curve), (ii) the effective action $S^{(1)}_{\rm eff}$ including the lowest order corrections in χ (dark purple curve), and (iii) the complete effective action where the χ field has been integrated out exactly (light purple curve).

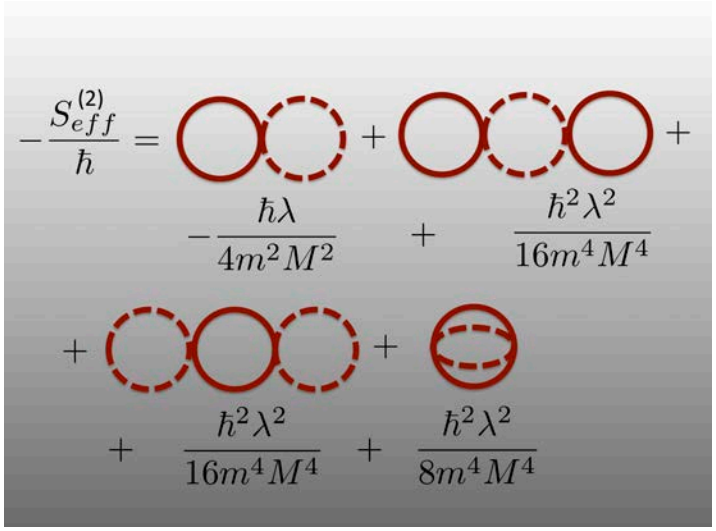


Figure III.4.23: *Effective action expansion.* The diagrammatic expansion of the effective action for the toy model of equation (III.4.26), and the expression for the terms up to order λ^2 and $\hbar^2$.

,

field variable and extract an effective action $S_{\rm eff}(\varphi)$ for the φ field through the defining relation:

$$e^{-S_{\rm eff}(\varphi)/\hbar} = \int e^{-S(\varphi,\chi)/\hbar}\, d\chi\,. \qquad \text{(III.4.27)}$$

Is this very complicated you may ask? The answer is: if you have real space-time dependent fields it is quite involved, but in our little kindergarten theory, there are no evil agents that could spoil our curiosity. As you know the action function is just quadratic in ϕ as well as χ which means that the complicated looking formula involves just one Gaussian integral over χ:

$$\int_{-\infty}^{+\infty} e^{-a\chi^2}\, d\chi = \sqrt{\frac{\pi}{a}} = e^{-\frac{1}{2}\ln\frac{a}{\pi}}\,. \qquad \text{(III.4.28)}$$

In our integral we have that $a = (M^2 + \lambda\varphi^2/2)\hbar$ and we obtain for the integral $\left(\frac{2\pi\hbar}{M^2+\lambda\varphi^2/2}\right)^{\frac{1}{2}}$. So the effective action for the φ field becomes:

$$S_{\rm eff}(\varphi) = \frac{m^2}{2}\varphi^2 + \frac{\hbar}{2}\ln(1+\frac{\lambda\varphi^2}{2M^2}) + \frac{\hbar}{2}\ln\frac{M^2}{2\pi\hbar}$$

This logarithm $\ln(1+b)$ can for small b be expanded in a power series $\ln(1+b) = b - \frac{1}{2}b^2 + \frac{1}{3}b^3 + \ldots$.

Now on a quantum level we are supposed to draw all possible vacuum to vacuum diagrams: these are diagrams without incoming or outgoing lines. Are such diagrams possible? Well, yes, of course! We have drawn the first few diagrams in Figure III.4.23, where we listed them in powers of the coupling constant λ and included all diagrams up to second order. Applying the Feynman rules given in Figure III.4.21, we can in principle write down the amplitudes, but we are in particular interested in the coefficients of the successive terms and the powers in terms of the fields. After some algebra you get a result for the sum that is not too

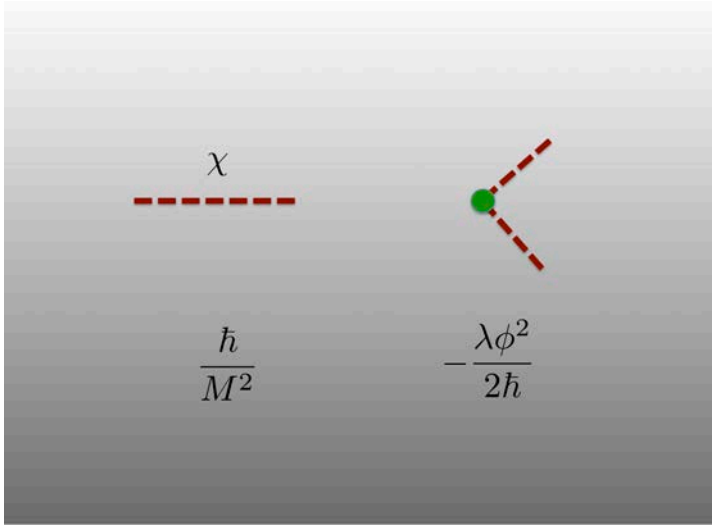


Figure III.4.24: *Effective Feynman rules for toy model.* These are the effective Feynman rules for the χ field if we treat the φ field as an external source corresponding to the green dot.

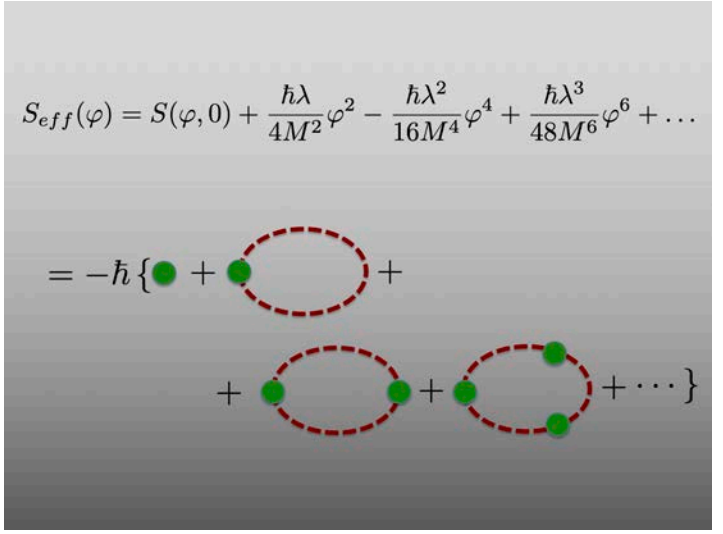


Figure III.4.25: *Effective action for φ field* This is the effective action for the φ field if we integrate out the high mass χ field. All diagrams have one loop and thus one power of $\hbar$, and they all contribute to the lowest order quantum corrections. The power of the coupling parameter $\lambda/2M^2$ is given by the number of propagators in each diagram.

surprising either:

$$S_{\rm eff}(\varphi) = S(\varphi,0) + \frac{\hbar\lambda}{4M^2}\varphi^2 - \frac{\hbar\lambda^2}{16M^4}\varphi^4 + \frac{\hbar\lambda^3}{48M^6}\varphi^6 + \ldots \tag{III.4.29}$$

This is an expression worth contemplating, because it exhibits many structural features of what quantum corrections on classical physics look like. We make the following observations:
(i) First of all note that the correction have a factor $\hbar$ so they vanish in the classical limit, in the classical limit the presence of the χ field decouples and it does not affect the effective φ theory.
(ii) The second remarkable fact is that summing over all χ contributions, which is what integrating the field out means, generates self-interactions of order n with an effective coupling constant $\lambda_n \sim \hbar(\lambda/M^2)^n$. Most important is the lowest order term quadratic in φ: in other words, it will shift the mass to $m^2_{\rm eff} = m + \frac{\hbar\lambda}{2M^2}$. The take home message is, that quantum corrections may introduce novel interaction terms that were not there on a classical level.
(iii) One might also derive effective Feynman diagrams for the φ field where the higher order terms are represented as new couplings λ_{2n} labeling the strength of the vertices with $2n$ external lines. This is depicted in Figure III.4.26.
(iv) A fundamental question that remains at this point is whether the effective quantum theory can produce interaction terms that violate the symmetries (and therefore conservation laws) of classical theory. We see an example in the toy model above. The effective potential for the φ field has positive coefficients for the φ^2 and φ^6 terms but a negative coefficient for the φ^4 term, which means that the potential will have local minima at for $\varphi = 0$ but also for $\varphi \neq 0$. It would correspond to a metastable state where the mirror symmetry $\varphi \to -\varphi$ is violated. We briefly return to this question shortly when we talk about anomalies.■ ■

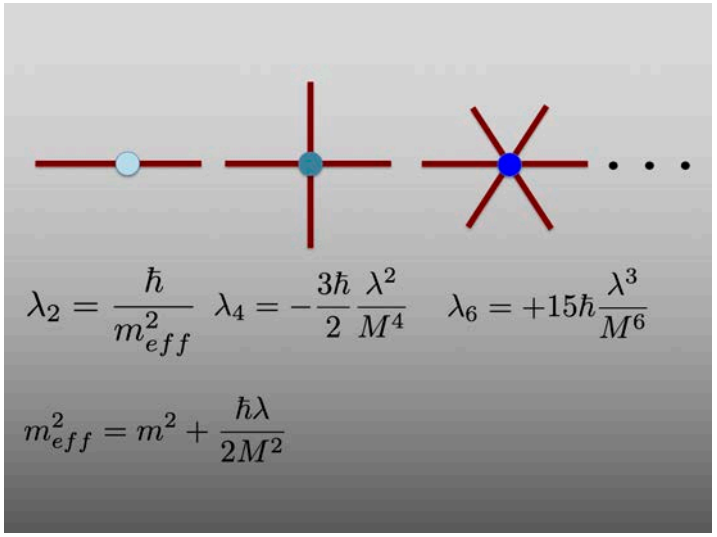


Figure III.4.26: *Effective Feynman rules.* The Feynman rules for the effective action for the φ field. The terms are defined as $\frac{1}{2n!}\lambda_{2n}\varphi^{2n}$ and yield the diagrams as shown.

Quantum fluctuations in QED

We pointed out before that we may specify a theory by postulating a set of fields representing the basic constituents and their interactions by giving the coupled equations they have to satisfy or equivalently by giving an energy or action function(al) in terms of them including the interactions. Such a model is characterized by its particular functional form which contains a number of parameters like coupling strengths and masses. These parameters are just the coefficients of the various terms in the action. Of course there are also the universal parameters such as the velocity of light and Planck's constant, which are hidden as we have 'set' them equal to one.

Now you would think that the parameters are directly determined by making measurements of them. Here we have to be careful because the story is not so simple.

In quantum theories even in the most idealized situations one has to deal with the effect of *quantum fluctuations*, because such fluctuations are an inevitable ingredient as a consequence of the uncertainty relations between position and momentum and time end energy. The size of the energy fluctuations grows inversely proportional with the spatial scale one chooses to look at. So, the theory describes also what the fluctuations are in these quantities and if one goes to smaller distances or higher momenta the effect of these fluctuations is that they will lead to significant differences between the *bare values* of the parameters that I wrote down in the equations and those that would effectively be observed. The parameters are indeed external but they are in fact corrected by the quantum processes described by the theory.

To make a consistent comparison with experimental results one should first calculate, then include these 'quantum' corrections and then choose the bare parameters in such a way that the observed data match the calculated parameters *including* the corrections. It's like buying a box of chocolates, since there may be a significant difference between the weight of the box as a whole and the net weight of the chocolates, as the wrapping may be surprisingly elaborate. The lore is that the more exquisite the chocolates the more elaborate the wrapping. Reality is similarly hidden from us by an elaborate quantum wrapping.

The calculations of these corrections turn out to be quite involved. What we like to do here is not so much doing such calculations as outlining the structure of what they involve. And what all that has to do with the scale dependence of the theory. Briefly stated: if one naively calculates these quantum corrections using the diagrammatic approach of Feynman, one finds that the calculations diverge, that they give infinite answers. This is not so much an indication that things are wrong, but rather that they are more subtle than you would naively expect. And Nature is subtle for sure.

What happens is that if one calculates the effect of certain quantum degrees of freedom these cause infinite changes in the effective parameters of the theory and that would render the theory useless, except when these divergencies can be 'subtracted' in a meaningful and consistent way that allows for a set of uniquely defined finite parameters after all. This procedure for dealing in a physically sensible way with these unwanted infinities is called *renormalization* which in turn can be understood in a more general approach called the *renomalization group*.

What one learns is that the renormalization procedure imposes serious constraints on the set of couplings or interaction terms one starts off with. Theories satisfying these constraints are called renormalizable and you will not be surprised to hear that the Standard Model of elementary particles and their interactions is a renormalizable gauge theory. However, Einstein's theory of general relativity is not renormalizable in the above sense, and the construction of a quantum theory of gravity is still best described as 'work in progress.'

Renormalization. Renormalization amounts to systematically extracting the finite quantum corrections to the parameters of the bare (classical) theory. It involves a rather technical two-step procedure to handle the infinities that pop up in the calculations of quantum corrections to masses and other coupling constants. The first part is *regularization* of the divergent expressions. This can be done in many different ways, but the simplest conceivable is to just introduce a cut-off in momentum space. This means that we simply ignore the contributions of very high momentum fluctuations. The second part is to introduce a *subtraction* depending on the cutoff, which renders the calculated amplitudes finite. The subtraction involves the introduction of *counter terms* in the action, and once these have been introduced one can take the limit of the cut-off to infinity. The dependence on the cut-off has disappeared and one is left with a finite physically meaningful result

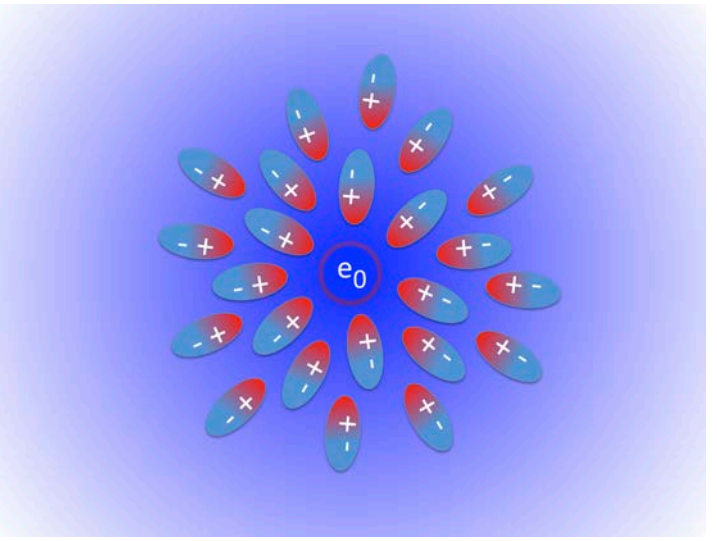


Figure III.4.27: *Virtual electron-positron pairs.* Vacuum fluctuations in the electromagnetic field give rise to a cloud of virtual electron-positron pairs that effectively screen the 'bare' charge and make the effective charge distance or momentum dependent.

In practice the contribution of the quantum fluctuations depends on two things: (i) a momentum cut-off Λ which indicates that one only takes into account fluctuations larger then a certain spatial scale $d \gtrsim 1/\Lambda$, and (ii) on how accurate one calculates the effect of the fluctuations on the parameter values of interest. The calculated parameter change is encoded in what is called the β function, and this function can be calculated to an increasing degree of accuracy. We have discussed already examples which showed that these technical considerations are crucial in determining in which parameter domains one may expect results that do or do not make sense. In the following paragraphs we will give some remarkable results that will show the analytic power of these methods if it comes to understanding the asymptotic (high-energy) behavior of physical systems and the theories that describe them.

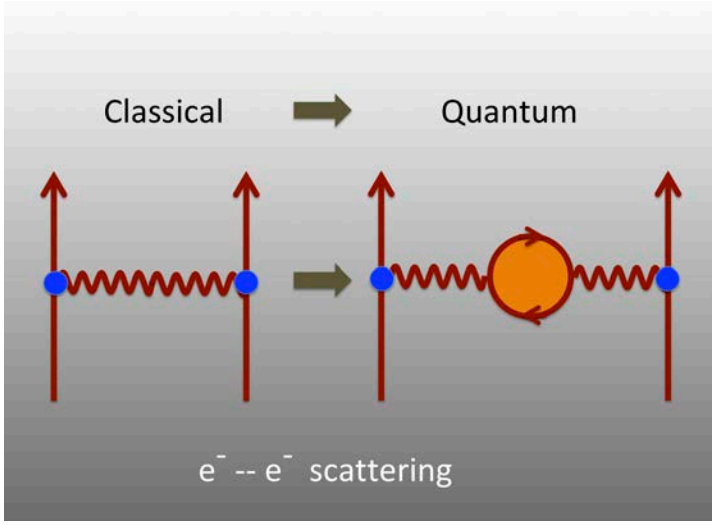


Figure III.4.28: *From classical to quantum process.* The Feynman diagram on the left represents two electrons that scatter of each other by exchanging a single photon. This diagram yields the classical result. On the right we give the quantum process where we have 'dressed' the photon propagator with a 'blob' which means that all quantum corrections have been included.

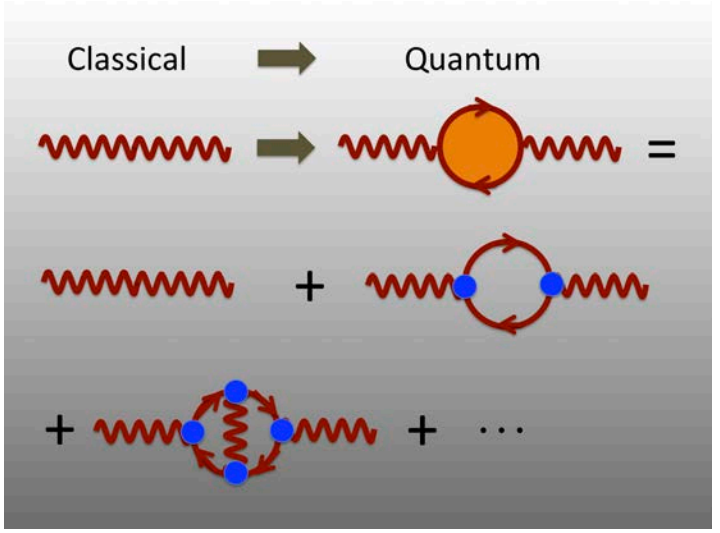


Figure III.4.29: *Quantum corrections.* Corrections due to virtual processes to the photon propagator. The blob has a systematic expansion in terms of ever more complex diagrams. Two blue dots lead to a factor $\alpha = e^2/4\pi\hbar c \simeq 1/137$ in the contribution of the diagram to the quantum mechanical scattering amplitude. So higher order terms (in α) become smaller and usually we include diagrams up to second order.

A realistic example: Vacuum polarization

We think of force laws like the gravitational law of Newton or the Coulomb law of electrostatics as specifying the strength of a force depending on some charge or mass and depending on some variable like the distance and then there is also an interaction strength, which is a dimension full constant (parameter) to be determined through experiment. That means we have to measure it at some characteristic scale and then assume it is constant not only in time but also in space. Both assumptions may be challenged. It may well be that by going to smaller or larger distances the effective coupling constants if one measures them would change.

Let me indicate why the effective coupling might change by exploiting some of the intuitive notions we have mentioned before. It is clear that if we have a charge for example, the field around this charge will become ever stronger at small distances. The energy density of the field increases and may at a certain distance become so big that it becomes possible by Einstein's $\mathrm{E} = \mathrm{mc}^2$ law, that charged particle antiparticle pairs are created near the charge. The idea is that the 'empty' space is not empty at all but filled with electron-positron pairs that form a cloud around the charge. This cloud will in fact screen the 'bare' charge of the electron. This means that at a distance further out we see an effective charge that will be smaller than the charge we started off with. Translated in the language of the coupling strength of the charge to the field, we see that it is not constant but depends on the scale at which it is measured. The amount of screening depends on at what distance we look at the charge. We say that the vacuum becomes polarized. As we measure at some distance it is interesting to ask whether we can find out what the bare charge

would be, or what the effective charge at other distances would be. The coupling constants may be on the run but where are they going? Is it going to be ever smaller or ever bigger and maybe become infinite? In theories of various sorts describing different types of interactions many different scenarios present themselves including the possibility that bare parameters that were chosen to be zero become non-zero due to these quantum fluctuation effects, which basically amounts to saying that the theory itself acquires new extra parameters that you didn't put in at the start! Under certain circumstances theories are apparently capable to 'improve' upon themselves. One might say that taking this kind of background into account provides insight in the range of validity of the theory one started off with and that is certainly a remarkable conclusion that deserves a closer look.

A divergent diagram. Let us consider the two-point function for the photon. In the second line of Figure III.4.29 we have drawn some Feynman diagrams that describe processes that contribute to the propagator or two-point function for the photon. In the first, the wiggly line is just the bare propagator which in momentum space is just given by the expression:

$$S(k) \simeq \frac{1}{k^2}$$

So this describes a mode with momentum of the photon propagating between two space-time points. The second diagram with two interactions where an electron-positron pair is created and subsequently annihilated. It is a so-called *virtual process* because there are no external lines connected to the closed fermionic loop. Now momentum is conserved in the interaction points so overall that means that the momentum carried by the ingoing photon must be the same as that carried by the outgoing photon, and at the vertex it implies that if the electron created has momentum p, then the positron has to have momentum $k - p$. If we just do the counting of powers of p, the propagator of the electron yields a factor $1/(p-m)$, and the positron a factor $1/(k + p - m)$. The problem arises because we have to sum or integrate all possible amplitudes, which means all possible values of the momentum p going around the loop. so we have to calculate an integral

$$\int \frac{1}{(p-m)(k+p-m)}\, p^3 dp \simeq \Lambda^2$$

For large p the dominant contribution comes from

$$\int p^3/p^2\, dp = \int p\, dp = \infty\,.$$

In other words, the integral behaves badly and is divergent! This is bad news because we know that physical amplitudes and probabilities are finite. What we need is a way to manage the deluge of infinities popping up in our calculations in such a way that physically meaningful results are obtained The infinities have to be artefacts of our calculational methodology otherwise the theory makes no sense.

This leads to the intricate protocol called *renormalization* that we have mentioned before. It refers to the three step procedure, where we first regulate the divergencies, then *subtract* the would be divergencies, which allows to redefine or renormalize the fields and parameters in the theory in a consistent and unique way.

Regularization and renormalization. The first step we take is to in some way *regulate* the divergent integral by introducing a high momentum cut-off, meaning that we limit the momentum range we integrate such that $p \leqslant \Lambda$. Then the leading term will be quadratic in Λ as indicated in the equation above.

Once you have applied such a *regularization* to all the divergent expressions, *renormalization* means that you apply a well-defined procedure to *subtract* the divergent expressions in a consistent way that leaves you with unique finite results for the quantum corrections to any diagram with given external lines. However, there are only a finite number of renomalizations (correction factors) you can im-

plement in a given field theory; you can basically renormalize the fields, the masses and the coupling constants and that's it. So for electrodynamics you could at most accommodate two field, one mass, and one coupling constant renormalization. There are dependencies between them and one is left with three correction factors, Z_1, Z_2 and Z_3 associated with the renormalization of the wave function(s), the charge and the fermion mass respectively. The fact that QED is *renomalizable* means that if you calculate all diagrams for all amplitudes to arbitrary order in the coupling constant, all divergencies that you will ever encounter can be absorbed in those three constants. This is by no means obvious; it means that the theory has to meet certain exquisite requirements We will have more to say about what that means and which generic properties determine whether a theory is renormalizable or not.

Let us reflect for a moment on what the above technical rather magical manipulations have to do with the main subject of this chapter which is 'scaling' and 'scale invariance.' It is quite clear that once you introduce a cut-off or any other way to regularize the theory, then that will break any form of scale invariance, precisely because we explicitly introduce a scale in the theory 'by hand.' And though the results claim to be independent of the particular value of the cut-off, renormalization nevertheless deeply affects the high-energy asymptotic behavior of quantum field theories and in particular spoils the scale invariance one might have expected.

The cut-off and the subtraction point

The role of the cut-off is rather profound. With a bit of common sense one would say: of course there ought to be a cutoff because the theory may not be fit to describe fluctuations in the medium below a certain scale. Think of a fluid which on a macroscopic level is a continuum, but if we go down in scale we know that it is ultimately just a collection of molecules and on that scale the continuum assumption is certainly a bad one. Evidently in such a case it is the interatomic separation in the liquid that sets the scale for the distance cutoff $\mathrm{d} \sim 1/\Lambda$. Let us now turn to the all-important question of the accuracy of the β functions, i.e. the functions that describe the scale dependence of the effective parameters in the model. The arguments became rather subtle to a point where even the scientist themselves became utterly surprised by the success of their calculations. What happened? In many cases the difference between the measured quantities and the calculated ones grew ever larger with increasing momentum. And indeed new parameters had to be introduced in the bare energy function. What one did was to just introduced so called counter terms also depending on the cutoff introduced that cancelled the calculated effect and after that let the cutoff go to infinity (or zero), so that the difference ended up being finite and independent of the cutoff. The physicists developed a well-defined procedure, or maybe we should call it a calculational trick, called *renormalization* that would lead to predictions free of ambiguities, if and only if after some given order in the approximation scheme of the beta function no new parameters had to be introduced. That means that after a certain point the number of parameters of the theory would stay fixed and finite. Renormalization would then only change those parameters, and that was considered admissible from a physical point of view, though mathematically one was kind of jiggling infinities to fabricate finite numbers that should fit the experimental data.

But as usual the proof was in eating the sausage without advertising too much what went in it. And the results turned out to be splendid and the renormalization methods allowed us to calculate many new physical effects with exceptional precision. For example, the pinnacle of such calculations is the high order calculation of the anomalous magnetic moment of the electron which matches experiment up to 11 significant digits! Now that is what one calls hard science!

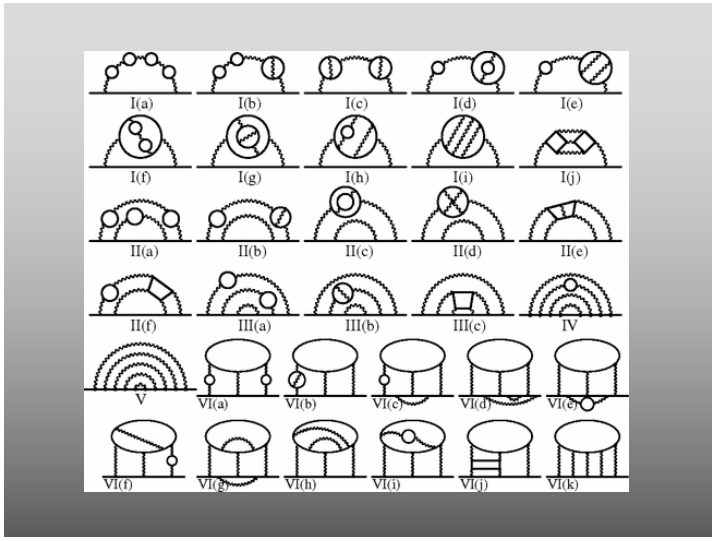


Figure III.4.30: $g-2$ *diagrams.* Some of the tenth order diagrams contributing to the calculation of the anomalous magnetic moment of the electron or muon. (Physical Review Letters109.111807, 2012)

In 1987 the experimental measurements (R. S. Van Dyck, Jr., P. B. Schwinberg and H. G. Dehmelt) reached the unbelievable precision:

$$a_e = (g-2)/2 = 1159652188.4(4.3) \times 10^{-12}\,.$$

The heroic QED calculation to the tenth-order in perturbation theory involving 12,672 diagrams performed by the Japanese team of Aoyama, Hayakawa, Kinoshita, and Nio produced the theoretical value:

$$a_e(\mathrm{theory}) = 1159652181.78(77) \times 10^{?12}\,,$$

which was published in 2012. To give you an idea of what this looks like we present some of the tenth-order diagrams in Figure III.4.30.

Anomalies. If regularization violates the symmetries of the classical action, we produce anomalies. The would-be conserved current is no longer conserved, the divergence of the current is no longer zero but there will be an anomalous source term in the quantum version of that law. So the question is how serious that is. What it means that in the quantum real world we would see processes that violate some naively expected conservation laws. For example, there is a famous decay of a neutral pion π_0 into two photons the would be forbidden but actually has been observed, so such anomalous processes do occur.

Now there is one important restriction here, as we have argued, gauge symmetries lead to electric or color charge conservation and it is known that if we break local gauge symmetries, that leads to severe inconsistencies and the theory would become non-renormalizable. So, in the first place we have to make sure we have a gauge invariant regulator. However, that may not be enough, and one must make sure to adjust the particle content of the theory such that the contributions of the different particle species to the anomaly cancel. This has indeed led to the constraint of the *family structure* of the Standard Model. If the particles appear in what we called 'families' than the cancellation of all gauge anomalies is guaranteed.

As a matter of fact here again it is the gravitational interaction which is after all a gauge theory which has a gravitational anomaly, which makes the 'naive' perturbative quantization of Einstein's general theory of relativity a well-established night mare! In fact it is exactly why an anomaly free gravity theory pops up in string theory. It turns out that the gravitational anomalies cancel in ten-dimensional space-time, where strings supposedly live.

Further reading.
On scaling, renormalization and critical phenomena:

- *Fractals*
  John P. Briggs
  Touchstone Books (1992)
- *Quantum Field Theory*
  David Skinner
  Cambridge University (Lecture notes)
- *An Introduction To Quantum Field Theory*
  Michael E. Peskin and Daniel V. Schroeder
  CRC Press (1995)
- *The Theory of Critical Phenomena: An Introduction to the Renormalization Group*
  J.J. Binney, N.J. Dowrick, A.J. Fisher and M.E.J. Newman
  Clarendon Press (1992)
- *Phase Transitions and Renormalization Group*
  Jean Zinn-Justin
  Oxford University Press (2013)

Complementary reading:

- *The Fractal Geometry of Nature*
  Benoit Mandelbrot
  W. H. Freeman and Co. (1982)
- *Fractals: Endlessly Repeated Geometric Figures*
  H. Lauwerier
  Princeton University Press (1991)
- *M.C. Escher: Art and Science*
  H.S.M. Coxeter, M. Emmer, R. Penrose and M.J. Teuber Eds
  North-Holland (1986)
- The Mathematical Side of M.C. Escher
  Doris Schattschneider Article in Notices of the AMS, Volume 57 nr 6 (2010)
- *Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies*
  Geoffrey B. West
  Penguin Press (2017)

Nature in search of itself.

Science is a deeply human endeavor, as it requires the unique combination of basic capabilities like curiosity, reason, intuition, creativity and collaboration. It expresses the collective curiosity of mankind and has resulted in the double helix of science and technology that keeps transforming our world over and again. It embodies a cumulative, evolutionary process that continuously creates new options for society while at the same time forcing it to face the severe ethical dilemmas that come along.

All of us have witnessed how science has profoundly affected the human condition and transformed society, and how in many instances it managed to transcend man's painful political, ethnic, and religious differences. As such it is a true cornerstone of civilization. At least as long as we can ensure that it does not fall prey to all kinds of abuse by dark forces bent on power and financial or political gain only.

If knowledge is our destiny, then that feeds the hope for carving out a gateway to a common, global understanding of the world and our options for governing it. It could lead the way towards an inhabitable future for all of us.

Chapter III.5

Power of the invisible

> Im ganzen habe ich jedenfalls erreicht, was ich erreichen wollte. Man sage nicht, es wäre der Mühe nicht wert gewesen. Im übrigen will ich keines Menschen Urteil, ich will nur Kenntnisse verbreiten, ich berichte nur, auch Ihnen, hohe Herren von der Akademie, habe ich nur berichtet.
>
> *Franz Kafka*, in *Bericht fur eine Akademie* [1]

In this concluding chapter we briefly recapitulate our journey through the quantum wonderland. It is a kind of mirror image of the introduction. The difference is that with the knowledge we have acquired along the way there is more room to reflect on the places we visited. This also means that there is some room for more subjective statements.

A pillar of wisdom? The cartoon on the right by Pete Ryan appeared in the *New York Times*. For me it is an ironic pillar of wisdom depicting not only the wisdom itself, but also our winding roads towards it. That process starts in quite an orderly way at the bottom with a number of parallel strands going straight up. At some point you start wondering why the strands go up so perfectly straight and parallel. And as soon as you start to question the given narrative things start to diverge. The lines start wiggling and before you know you are caught up in a huge entanglement, a huge confusion, a spaghetti like mess of doubt and contradiction. How to move forward? How to get out of this mess? And yes, every time, as by some miracle you manage to surmount the problems and look what

Figure III.5.1: A modern pillar of wisdom? (Source: Pete Ryan, NYT, Jan. 7, 2022)

[1] On the whole, at any rate, I have achieved what I set out to achieve. But do not tell me that it was not worth the trouble. In any case I am not appealing for any man's verdict, I am only imparting knowledge, I am only making a report. To you also, honored Members of the Academy, I have only made a report. (translation: Willa and Edwin Muir)

happens: things come together again, and they coalesce into a new perception of reality symbolized by the beautifully ornamented capital. Like a crown on your labor. You managed to beat the minotaur hidden in that labyrinth of strands.

I suppose the artist has forgotten about the many dead-end streets that are also part of the tangle. Maybe the artist was inspired by the path integral approach to wisdom where only paths from A all the way to B have to be included. In fact, *all* of them have to be included, not just the shortest or the most beautiful, but also the less obvious, maybe obscure and low probability routes. Anyway, after working your way through all the possible paths you are bound to end up at a next level of knowledge and understanding. Yet another shoulder of a giant to stand on, yet another step on Cantor's devil's staircase to ultimate knowledge and may be wisdom....

Summary and outlook

To see the power of the invisible in a way that supersedes blind faith, it must be made discernible first. And that is what empirical science is about, inventing the observational tools that allow us to see those things that have always been there, but were hidden from the naked human eye. Why worrying about the invisible, you may ask, as long as the visible suffices to keep us busy and to fully occupy our fragile minds? Curiosity to know what may or may not be beyond what we can see is one of the ultimate drivers of our existence, of discovery, and in the long run of understanding, reason and survival.

Who am I? If you would ask me who I really am, I may start by telling a nice story, probably a dressed up CV of some sort centered about my major accomplishments. In certain cases I may even disclose some personal details. And if you keep pushing me, it may turn into a narrative about my childhood, my family and its traditions. And by talking about family treats I have, without mentioning, entered the realms of heredity and of genetics. The narrative loses some its ultra personal features and turns into a more generic, though still fully anthropocentric, perspective. I will for example not mention that features like my sense of humor, or need to physically be in touch, or my habits of impressing others, or getting enraged about futilities, probably go all the way back to my primate or for that matter rabbit-like ancestors.

You understand what I am driving at: the deeper I search myself and the world in which I live, the less personal the story becomes, the more abstract it will be, and the less it will refer to the plainly visible or the specifically human. If your interrogation were to go on indefinitely, I might just jot down some quantessential formulas in the end. And that is how the science of the invisible enters our conversations as a relevant resource of reliable knowledge, leaving the limitations of anthropocentricity behind. Maybe that is the power of the invisible.

The mission of physics. Physics is an empirical science which concerns the art of making discoveries through making ever more sophisticated observations. It wants to know what nature looks like and how it works on all scales. We have to admit that it certainly paid off when Galileo supposedly threw stones and wooden balls from the Pisa tower and carefully listened to them hitting the pavement! We make progress by building models and improving on them. The models are supposed to not just fit data but more in particular to explain the different patterns of data by relating them through causal relationships expressed through mathematical equations.

On all scales there is the question what the relevant degrees of freedom are, and to understand their behavior, like structure formation through binding or a particular dynamics, we need to understand the interactions between these relevant constituents. Dynamical processes are gen-

erated by interactions or forces between constituents and that makes their overall effect often hard to predict, exactly because coupling systems introduces feedback loops. This is a general feature that holds both on the classical and quantum level. It is particularly true for many particle systems, but as we have seen, it also holds for space-time. So, let us once more look at the quantessence at large.

This book's approach. This three-volume book is a somewhat experimental and ambiguous 'go in between' in the sense that it tries to interpolate between a 'laymen account' and a – I hate to say this – 'textbook' of a sort. Is it possible to go in-between without losing two audiences at once? My publisher will undoubtedly let me know immediately, I am sure!

Another question that crosses the mind is whether all those *Wikis* make books like these not obsolete? I think the answer is a firm 'no' and would claim the opposite. These books attempt to be more than a encyclopedia and give a coherent account of large range of topics that together form a huge subject in science. The aim is to provide a critical guidance for which items out of the small infinity of *Wikipedia* entries are relevant if you want to go quantum. I can only hope that these books did indeed give you an informed steer on when and where to go for additional *Wiki-wisdom*, and what the keywords were to look for.

It's the math, stupid! In confronting quantum realities this could be the analogue of the political maxim 'It's the economy, stupid!' that was coined by the American political analyst James Carville in 1992. He wanted to emphasize that even the most basic knowledge of economy would stop people from making absurd claims about everyday economic realities. We have used a lot of mathematical language mainly to keep the arguments transparent and unambiguous and to prevent us from committing crimes against logic. But we softened our approach by paraphrasing the math with lots of prose as to keep the story accessible. However, making that choice we sacrificed a principal asset of mathematics, namely, that it is extremely concise and allows you to make precise yet brief arguments. The true aesthetics of mathematics is deeply rooted in this idea of eliminating all the unnecessary. In that respect math is the opposite of show business: no window-dressing allowed. We exploited the unambiguous and transparent character of the mathematical formalism, but at the same time blurred its purity by – in parallel – talking extensively about what it means and using lots of illustrations. We immersed our math formulas in the 'unnecessary' to keep them accessible and part of the conversation. You could say that we fell back on show business after all.

The three track narrative. In an attempt to help overcome the common fear of formulas and keep the contents manageable I adhered to a storytelling philosophy where the narrative followed three tracks in parallel. The first was a pictorial one, as I included over 450 illustrations, the second was the rather extensive use of equations, and the third track consisted of extensive prose. The latter is there in its own right, but also to bridge the gaps between pictures and formulas. The interplay between these tracks hopefully allowed you to grasp this wonderful body of fundamental knowledge in the heart of science. I am convinced that it made you at least 'conversant' about the *quantessence* of things.

The *quantessence* in retrospect.

Let us look back at the three volumes that make up this quantum trilogy with Figures III.5.2 in mind. The reason why this trio has such a wide scope is the fact that quantum theory is a general set of principles that nature appears to obey on all scales, at least as far as we have been able to test. It applies to different types of systems, where the translation of the fundamental quantum principles get a different mathematical implementation and outlook.

Three volumes.

Volume I. In the first volume we have devoted quite a bit of time and room to provide a wide background by recalling the basic concepts of classical physics. This to provide a setting in which the quantessential parts of the subsequent volumes stand out more clearly.

In Chapter I.1 we briefly reviewed of the central achievements of classical physics. And in Chapter I.2 we extended that with the basics of relativity, geometry, and classical information theory.

Chapter I.3 looks at the universal constants of nature and what their meaning is. We showed how through dimensional analysis these constants set natural scales linked to certain classical and quantum phenomena.

In Chapter I.4 we descended the quantum ladder in a systematic way from the atomic scale down. This culminated in a description of the Standard Model for the elementary particles and the fundamental forces between them. We then continued with excursions into the speculative domains of supersymmetry and string theory as possible approaches to a consistent quantum theory that includes gravity: a quantum theory that would unify matter, radiation and space-time.

Volume II. In the second volume we introduced the mathematical framework and mostly applied it to basic systems like qubits, electron spins, particles and simple field theories.

In the Chapter II.1 we discussed concepts like the Hilbert space of states, a vector space where the linear superposition principle holds which quite directly leads to the possibility of *entangled states* which are uniquely quantum. These states lead to intriguing paradoxes like 'Schrödinger's cat' and the EPR paradox, but at the same time opened the possibility of quantum teleportation and quantum key distribution.

In Chapter II.2, we introduced the observables as operators acting on Hilbert space. This identification led to quantessential notions like the incompatibility of observables, which in turn give rise to the fundamental uncertainties as expressed by Heisenberg's uncertainty relations. We also went into various aspects of particle-wave duality, leading to particle interference phenomena as discussed in Chapter II.3.

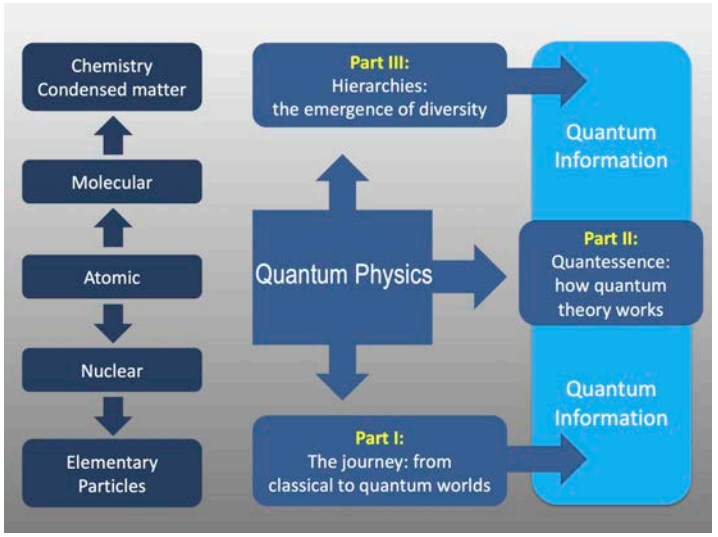
We demonstrated that the vastly different quantum setting allows for a new type of information processing and computing with a far-reaching technological potential. This is a major challenge and has become a high priority effort for the worldwide community of quantum condensed matter physicists. And in parallel to the struggle to produce scalable and reliable hardware there is now also a booming branch of quantum software developments.

In Chapter II.5 we explored a topological argument for the exclusion principle and the spin/statistics properties of quantum particles.

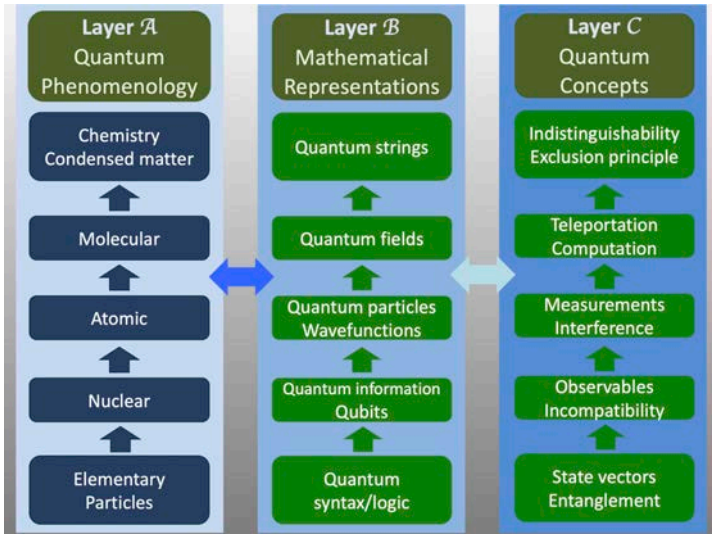
Symmetry considerations play a central role in all fields of modern physics and chemistry. We therefore concluded the second volume with a chapter entirely devoted to the meaning and quantum implementations of symmetry and its breaking.

Volume III. In the third volume we showed how the physics of the early cosmic evolution in an expanding and cooling universe is completely governed by the quantum laws. The resulting structural hierarchy of matter reflects how the various fundamental forces played dominant roles in successive stages of that evolution.

After discussing the basics of molecular (chemical) physics, we turned to the many-body physics of condensed



(a) The book at large.



(b) Quantum layers.

Figure III.5.2: *Book summary.* The quantessence in retrospect.

states of matter. First we described the types of order that the substrate of atoms or ions may exhibit, like crystal lattices of all sorts and their symmetries. We also considered the defects or imperfections that may form in such highly ordered states of matter. These defects often carry quantum numbers that are conserved for topological reasons.

In Chapter III.3 we turned to the electron collective and how that gives rise to many surprising quantum phenomena like various types of conductivity and magnetism, from semi- to superconductors, quantum Hall states etc. It is amazing to see how many novel states of matter are possible in the quantum regime.

We closed our quantum excursions with Chapter III.4 on the properties of scaling, first in the realms of geometry and then in context of dynamical systems. In the quantum regime leading the notion of *renormalization*, which boils down to a systematic scale dependent redefinition of the parameters that define the model. In quantum field theory this stands for sophisticated procedures of juggling with infinities leading to a state of peaceful coexistence with them, by producing unambiguous finite answers. In addition to the understanding of phenomena like the *confinement* of quarks, the *renomalization group* approach provided a powerful approach to critical phenomena in general.

Three layers.

Layer $\mathcal{A}$: Down and up the structural hierarchy. There is a subtle difference between the first column of the Figure III.5.2(a) referring to the Volumes and Figure III.5.2(b) referring to the layers. In the first figure the arrows are pointing downwards from the atomic scale to the scale of quarks and leptons, while in the second they are all pointing upwards. We recall that in Chapter I.4 we followed the quest for ever more fundamental building blocks of matter indeed following the arrows down. However, in Chap-

ter III.1 we did the opposite, and followed the time path, not of the human quest, but of the true cosmological history by showing how the hierarchy of matter starting from the Big Bang all the way up to the molecules of life came into being. This perspective was of course forced upon us after understanding the evolution of space-time according to the Big Bang scenario described by the theory of General Relativity.

Layer $\mathcal{B}$: The hierarchy of mathematical realizations. The mathematical realizations of the basic quantum principles are shown in the second column of the figure. Simply stated, the system one considers defines what the basic degrees of freedom or dynamical variables are. Given the Hamiltonian one may then define the operators for the 'coordinates' and conjugate 'momenta' and postulate their canonical commutation relations. The structure of the corresponding Hilbert space of quantum states then follows. If the system cannot be solved exactly which is mostly the case, one usually starts from the non-interacting system, and uses that as the starting point for a perturbative approach of the system with interactions.

What the middle column shows is that at the bottom of the hierarchy, the most elementary quantum system is in fact the *qubit* or the spin-1/2 degree of freedom, with its two-dimensional Hilbert space. This system was extensively analysed in Chapters II.1 and II.2.

One step up we have the framework of *quantum mechanics* for a single particle, typically in an external potential leading to an infinite-dimensional Hilbert space of normalizable wave functions. These notions were introduced in Chapter I.4 in the section on 'Atomic structure' and we repeatedly returned to this topic in the second volume, and in particular in Chapter II.5.

At the next level of generality, we include special relativity which forced us to move from quantum mechanics to the framework of *quantum field theory*. Here the fields and their conjugate field-momenta are the basic degrees of freedom, leading to the multi-particle Hilbert space. This framework centers around *field operators* that allow for the creation and annihilation of particles and therefore allows for the implementation of the famous equivalence relation $\mathrm{E} = \mathrm{mc}^2$, for example as we see it in processes like pair creation and annihilation in QED. Field theory is the language of the Standard Model, but also for most of condensed matter physics. Quantum field theory is introduced in Chapter I.4 in the context of the Standard Model. We returned to some of the formal aspects in Chapter II.5 and apply it to the electron collective in Chapter III.3. Finally, the scaling and renormalization aspects of field theory were discussed in Chapter III.4.

A yet more general framework would allow for the consistent inclusion of general relativity: in other words the inclusion of the gravitational force implying the quantization of space-time itself. This mission is not completed yet. The most advanced models of this type are the *superstring theories* which we described towards the end of Chapter I.4. In this framework each string mode corresponds to a different quantum field. The string idea therefore unifies all fields and thus all particle types into a single theory. This theory has certainly deepened our understanding of the quantum properties of gravity, like black holes and resolved some of the outstanding paradoxes, but the theory has not yet led to unique explanations of observed phenomena like dark energy. And the predictions it does make, like the 10-dimensional structure of space-time, or the existence of a myriad of super particles, have not (yet) been confirmed by experiment.

Layer $\mathcal{C}$: Quantum concepts and their meaning. The third layer shows how the mathematically consistent framework raised a number of conceptual issues physics had to face. These issues concern the question of how to interpret the core of physical reality. The subtitle of the book is 'The *quantessence* of reality' because that quantessence has been shaking the foundations of many of our cher-

ished beliefs about what seemed to be self-evident features of reality, features reflecting our classical intuitions. These intuitions concern what the properties of physical systems were supposed to be, and what the role causality and predictability in their mathematical framing amounted to. What we have learned in a century of quantum developments is that these changes are radical and will be long lasting.

Starting at the bottom of the third column of Figure III.5.2(b), we see the that the structure of the space of states of any quantum system is a vector (Hilbert) space, meaning that the superposition principle holds, and that physical states are represented by normalized vectors. If we combine subsystems the total Hilbert space becomes to the (tensor) product space, implying that the dimension of the total space is the product of the dimensions of the subspaces. This structure implies the existence of *entangled states*, which are states that correspond to normed vectors in the total space that are *not factorizable*, that *do not* correspond to a direct product of two vectors in the subsystems.

Entanglement allows for the possibility of strong, very quantessential, instantaneous correlations between outcomes of measurements separated by arbitrary large distances. This led to a profound debate often referred to as the *Bohr-Einstein* debate about the locality and causality properties of physical reality. Experiments like the GHZ experiment that we discussed in Chapter II.4 convincingly showed the quantum interpretation to be correct.

Moving one step up in the column we mention that the mathematical structure of quantum mechanics implies that observables should be interpreted as (bounded) operators acting on vectors in Hilbert space. These should be thought of as (finite or infinite) matrices or differential operators which by acting will in general change the state. The fact that observables are no longer real-number-valued variables like in classical physics immediately leads to the problem of what a measurement exactly means. In the Copenhagen interpretation it means that the measurement outcome is a probabilistic one and furthermore that the act of measurement will generically change the state of the system. There is no longer a strict separation between object and subject when observations are made. We can no longer predict precisely what happens but can only calculate the odds. This in turn means that we leave the notion of classical determinism behind. Quantum means indeterminism.

Another important consequence of the fact that observables are operators is that they do not necessarily commute. The outcome of their successive action on a given vector may depend on the order in which you apply them. If the operators do not commute, the corresponding observables are called *incompatible*. This incompatibility lies at the root of the intrinsic quantum uncertainties in measurement outcomes so beautifully encoded in *Heisenberg's uncertainty relations*.

The structure of quantum reality also implies that we cannot copy a quantum state while keeping the original, this is known as the *no-cloning theorem*. However, what *is* possible is to transfer a quantum state from one system to another, and because of the entanglement property this can in principle be done instantaneously over arbitrary large distances. This possibility of *quantum teleportation* turned the entanglement property into a blessing in disguise. It enables another level of cyber security in data transfer.

Further consequences of the quantessentials become clear from the information perspective. The quantum states allow for storage of information, and this led to the introduction of the *qubit* as the quantum analogue of the digital bit. Quantum mechanics allows for unheard possibilities to process this quantum information. We see all around us that a major quantum information revolution is on its way, a revolution that both on the hard and software side will radically transform our computational abilities.

A final radical ingredient of quantum reality manifests itself if one studies the collective behavior of many particle systems. First of all because particles of a certain type correspond to basic modes of a single quantum field, they are *indistinguishable*, they have a family name but no first name, so to speak. In addition, there is the possibility of *exclusion*, saying that there cannot be more than one particle in a given quantum state. This verdict is anchored in the quantum interpretation of the Dirac field. We addressed these fundamental properties of quantum particles in Chapter II.5 and linked them to the topological properties of the two-particle Hilbert space. Indistinguishability and exclusion each modify the statistical properties of many body systems and create entirely novel possibilities for the physical states of these systems. These possibilities have made quantum condensed matter physics into an inexhaustible source of technological innovations.

Altogether the beauty of the conceptual notions which surfaced in the third layer are a direct and therefore necessary consequence of the basic logical structure of quantum theory. There appears to be no way around them and more and more we start to appreciate how they enriched and broadened our perception of the roots of reality. They embody a true revolution in our understanding of the physical universe that found its translation into powerful new technologies that radically transformed our daily lives, and will keep doing so.

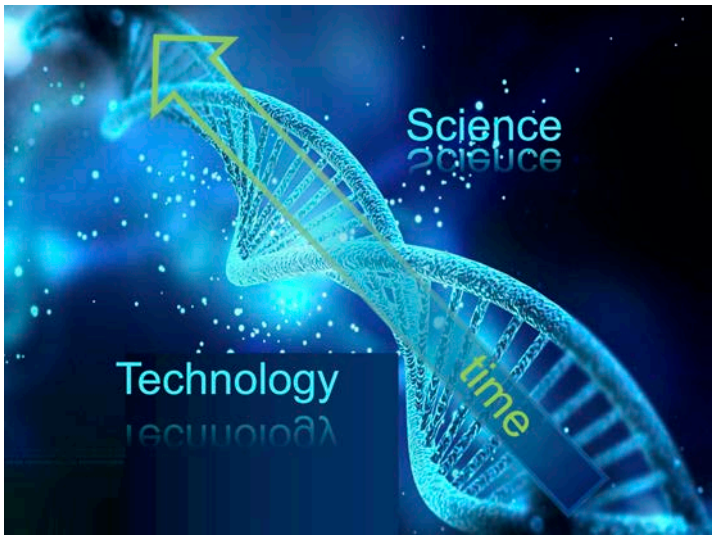
The many topics we didn't talk about. Many of the quantessential subjects we only touched upon superficially deserve chapters or books on their own. We spent a section on the miraculous properties of Carbon but what about a chapter on the virtues and technological blessings of silicon? What about the nano-sciences? What about an extensive review of an ever-growing list of alternative interpretations of quantum theory, like the 'many-worlds interpretation' proposed by the American physicist Hugh Everett in his doctoral thesis at Princeton University in 1957? Indeed, there are many topics which are relevant that I chose not to focus on and only mentioned in passing.

The main reason for these shortcomings is that I wanted to stay faithful to the subtitle of the book and focus on the *Quantessence*, the well-established fundamental aspects of the quantum reality. The perspective that shook the scientific world a century ago and lead to an unlimited extension of technological opportunities and realities that has by far not been exhausted or even been fully explored. As I emphasized all along, the era of quantum information technologies for example has only just started.

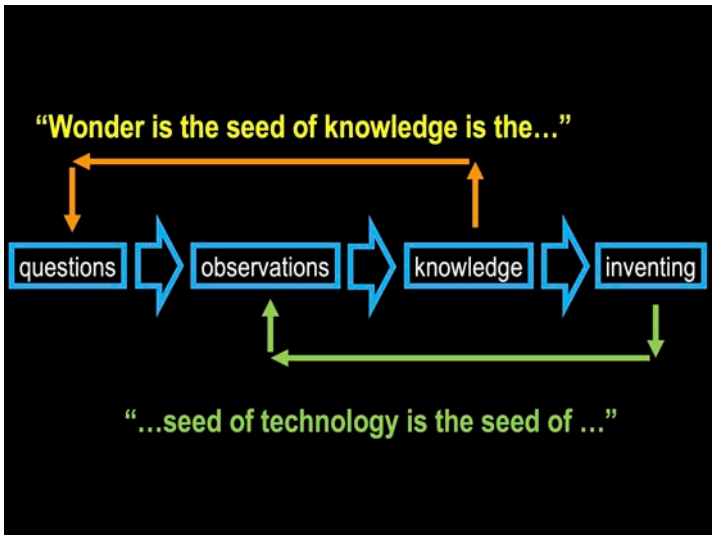
Common denominators.

The power of information as fundamental concept. Figure III.5.2(a) is just like the figure we presented in the Introduction to the book except that on the right we added a full column referring to the notion of information. It underscores that on all levels we may include an information science and computational perspective in the framework. All systems are in a sense information carriers and information processing devices, meaning that we set up paths with preset interactions between these carriers. Execution of a program or algorithm can be thought of as a particular class of dynamical processes. In this book we have repeatedly noted that the information science perspective involving algorithmic thinking is in an interesting way complementary to the more conventional theoretical physics approach involving calculus and differential equations, and it has led to surprising insights.

We encountered the notion of information towards the end of Chapter I.1 while introducing the notion of entropy as the logarithm of the number of micro-states corresponding to a given macro-state. It is a measure of information capacity of the system, or stated differently, for the information loss in going from the micro- to the macro-description of that system. It involves the aggregation of micro degrees



(a) Human evolution as driven by the double helix of science and technology.



(b) The positive feedback loop of science and technology producing knowledge, technology and the human expertise.

Figure III.5.3: *The double helix of science and technology.*

of freedom into far fewer macro degrees of freedom. In that sense entropy is a measure for hidden information. In Chapter I.2 we gave a small introduction to the basics of information theory as initiated by Turing and Shannon, and in the section on black holes we discussed the Bekenstein-Hawking entropy and the famous black hole information paradox.

In the quantum realm, we introduced in Chapter II.1 the idea of a 'bit mechanics' as the most basic of all dynamical systems leading to the notion of a qubit, with its two-dimensional Hilbert space. In the following chapters we illustrated many fundamental quantum concepts referring to this basic quantum system. In Chapter II.4 we talked about teleportation of quantum information, about quantum gates and circuits, and went into a rather detailed discussion of Shor's quantum factorization algorithm.

So indeed, the notion of information popped up everywhere justifying the blue column on the right-hand side of Figure III.5.2(a).

The power of symmetry as guiding principle. We saw that symmetry is a powerful notion with applications on all levels of the quantum ladder. This is reflected in the rich nomenclature involving symmetry concepts, like global versus local (gauged), space-time versus internal, exact versus approximate, and broken versus unbroken symmetry. It is not surprising that the notion of symmetry popped up in many chapters. We decided to devote Chapter II.6 to the many ways symmetry concepts have entered physics. In a sense it also deserves just like information a full column in Figure III.5.2(b).

Symmetries in classical as well as quantum physics are linked to conserved quantities.Therefore they lead to a transparent labeling of the physical properties of states. It allows us to give names to things like 'energy,' 'angular momentum,' 'charge,' or 'isospin.'

Symmetry considerations play a crucial role in analysing and understanding the solution spaces of the fundamental equations of quantum physics, like the spectra of single atoms and molecules as well as the states of many-body condensed matter systems. And symmetry served as a successful guiding principle in the uncovering of the underlying structure of subatomic physics encoded in the Standard Model as an expression of the underlying gauge symmetry.

Symmetry-breaking turned out to be a key concept to explain the many different guises in which symmetry manifests itself on all levels in nature. From Zeeman splitting on the atomic level to spontaneous magnetization or superconductivity on the macroscopic level, to the existence of the Higgs particle on the subnuclear level. Indeed, the idea of symmetry-breaking led to a unified understanding of the phase structure predicted by a wide variety of theoretical models.

The power of modelling as a discourse. Most models are quantitative in nature and by construction logically consistent. An ever-expanding body of symbolic relations that may be used to represent anything you can imagine. A human-made symbolic language ideally suited for a truly scientific discourse. Many of the great scientific turning points are cast in simple mathematical equations, or mathematically defined rules.

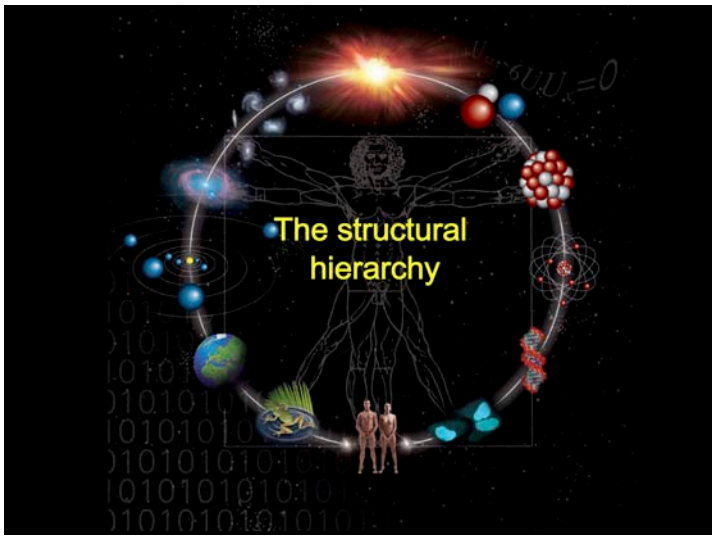
State-of-the-art modelling. Modelling is not only a way to talk *about* reality; it is also a way to talk *with* reality. It is a productive way of framing the scientific discourse. A state-of-the-art model is rarely completely correct. It has its strong and illuminating sides but also its weaknesses. So especially once the systems become complex with many hidden feedback loops and many coupling parameters one doesn't expect perfect predictions, and less so on the long-term future. What you gain in adaptability you lose in predictability. Think of modelling the climate or the spreading of viruses like Covid-19 or Ebola, or the endless efforts to properly model the good old economy.

The modelling activity furnishes a platform to study the effect of possible interventions. This is an interactive platform that can bring opposing interest groups together in a reasonable debate or negotiation, assuming both share enough purpose. Playing with the parameters of models gives a clear impression of what might go wrong, what the vulnerabilities of the system are, and what type of tipping points can occur. Models thereby can forge the highly needed compromises in order to be able to deal with the problems one is faced with.

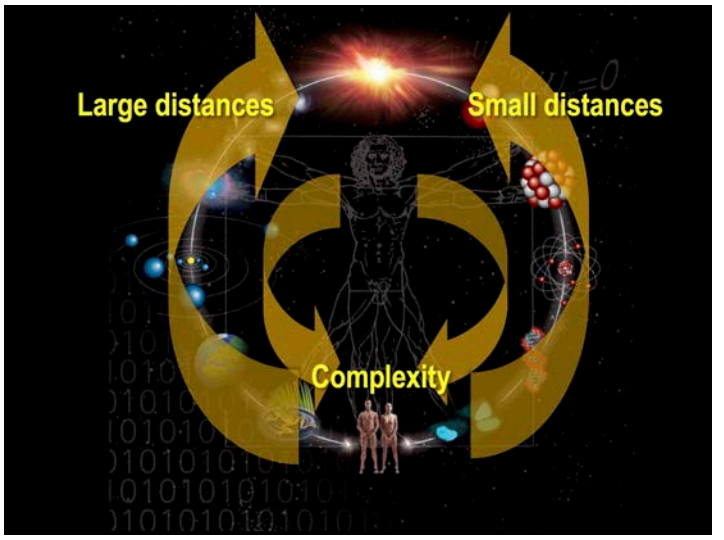
Analytic versus algorithmic thinking. We have stressed that a crucial aspect of scientific progress is the parallel development of mathematics as a language for modelling nature. Nowadays we should also include the crucial importance of computation and algorithmic thinking as powerful means to achieve progress in science. This concerns a wide range of methodologies, beginning with simple numerical methods to solve systems of mathematical equations to advanced simulation methods for complex systems like agent-based modelling. But also methodologies like machine learning to collect and analyse large data sets, algorithms to detect correlations, that make predictions possible without an actual understanding of the causal mechanisms underlying them.

Rule-based models. In this era of computational empowerment, we are increasingly driven away from completely analytical, closed systems of equations like those of Newton or Maxwell, to more evolutionary approaches like simple rule-based models. Rules that are iterated very, very many times and may lead to structural entities in which we recognize fundamental aspects of reality. This approach involves a shift from analytic to algorithmic thinking.

A key feature is that simple algorithms can generate extremely complex patterns with all kinds of emergent order. That emergent order is very hard to predict in ad-



(a) The structural hierarchy mapped onto a circle.



(b) Three fundamental frontiers.

Figure III.5.4: *The structural hierarchy of the material world and the basic frontiers of science.* In (a) we mapped the structural hierarchy onto a circle. Moving clockwise is moving towards larger scales, starting from 10^{-20} and extending all the way to 10^{+25} meters. The human scale is kind of in the middle. In (b) we indicated the three fundamental frontiers. On the left the large-scale frontier of astronomy pursued through space observatories like the Hubble and the James Webb. On the right the small-scale frontier of high-energy physics pursued at CERN and Fermilab for example. The arrows pointing towards the bottom symbolize the multiple frontiers of the life sciences including neuroscience. These naturally expand into the vast domain of information and computer science that are redefining the range and ambitions of the social sciences including economics.

vance using tools from standard analysis and geometry; its complexity can only be understood from actually running the algorithm for a sufficiently long time. We speak of *irreducible complexity* inherent to certain simple rule-based dynamical systems: for example, cellular automata or evolutionary pattern growth algorithms on networks, like John Conway's *Game of life.* The simplest way to find out what the structures are that emerge from a certain rule is to run the corresponding program long enough. We refer to the extensive literature on this subject by its pioneer and protagonist Stephen Wolfram who is also the founder and CEO of the successful software environment called Mathematica and Wolfram language. In his latest project aimed at 'finding a new fundamental theory of physics' he argues that all of quantum may be the product of iterating a simple rule-based algorithm! Another great mission, but for now also incomplete.

Scenarios for past and future

Science at large. In this final section I would like to put the whole quantum story in the wider context of science in general, a perspective that derives also from my earlier book titled *In Praise of Science: Curiosity, Understanding and Progress*. And in doing so I have adapted some of the imagery created for that book.

To me one of the most remarkable facts we are aware of is

that nature evolved from a random and structureless initial state with a uniformly distributed low information density, to a state of very high information content, very much localized in the most advanced of biological organisms such as human beings. It did so by following a set of strict rules we call 'laws of nature.' The most stupefying twist is that these rules have been hidden until we as human beings became aware of them after millennia of carefully researching and modelling what we observed. Indeed, nature seems to be in search of itself, becoming aware of itself through this concerted yet indefinite human effort.

The double helix of science and technology.

Let us focus a bit more on the mechanism underlying this process of progress as depicted in Figure III.5.3, On the left we see a schematic of what I have called 'the double helix of science and technology.' It is like a mutually inspiring, almost ritual dance, generating knowledge and technology, but also the expertise of scientists and engineers who are able to create and apply that knowledge. Paraphrasing Francis Bacon it visualizes the idea that 'wonder is the seed of knowledge' and 'knowledge is the seed of technology,' which in turn is the seed of new 'wonder' and scientific discovery.

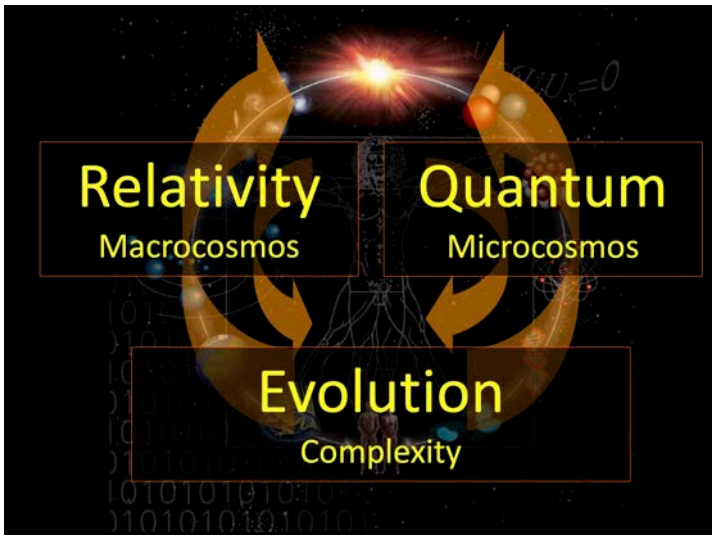
This perpetual machine works because technology also involves the invention of new instruments that shift the boundaries of what is observable. It pushes the observable in an objective sense. The domain of empirical investigation keeps expanding, generating an ever-growing body of knowledge! From instruments like microscopes and telescopes, all the way up to MRI machines, accelerators, and not to forget computers. The power to compute, to simulate numerically, as well as screening immense quantities of data for all kinds of correlations and patterns which are hidden from the human eye, is invaluable for human progress.

This human-made evolutionary process overtakes biological evolution in the sense that it continuously offers new options to humanity to move forward. I use the term options on purpose because it implies the notion of choice. The term progress suggests that society will always benefit, but that is not necessarily the case. What is certain, however, is that society will keep being bombarded with ethical and moral dilemmas, because those are inherent to that double helix of innovation.

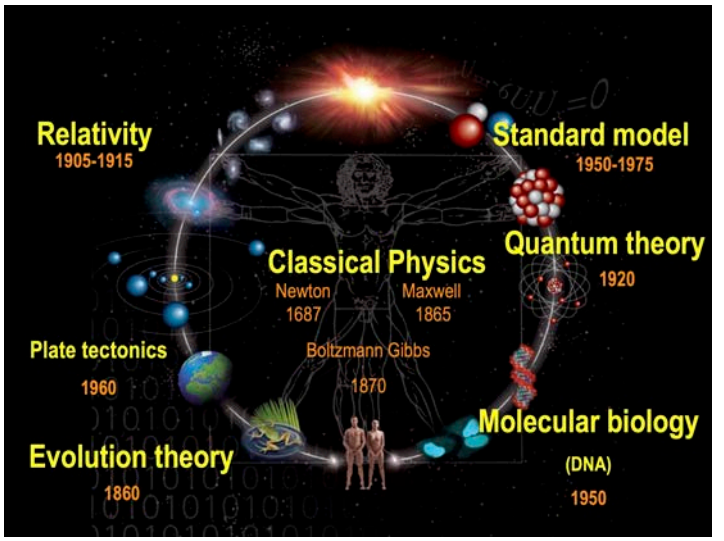
History has taught us that technology is a double-sided sword which may be used in constructive as well as destructive ways. And that means that it requires a society that has the ability to make the right choices and in particular manages to avoid a proliferation of the evil aspects of technological achievement. I think there is ample room for optimism but to close one's eyes for the risks and the dark sides that are certainly there, is dangerously naive.

Looking at the double helix of Figure III.5.3(b) one realizes that it is a magical machine that is not easy to stop. It is a positive feedback loop. It is hard to forbid curiosity or creativity by law but there have been regimes that did exactly that, a game only with losers. This machine is much more autonomous than most people are aware of. It takes a great deal of expertise and scientific awareness to navigate society in a way that the constructive opportunities get amplified, and the destructive ones are eliminated as far as possible. It is quite evident that good science does not work by popular vote. The scientific method is open to critique and rigorous analysis, but it is not democratic in the 'one man one vote' sense. That does not preclude that by the time new technological options present themselves to society one may hope that well-informed crowds will demonstrate their wisdom in governing their implementation.

This observation once more underscores the importance of fighting scientific illiteracy through broad educational programs introducing science and technology and raising the



(a) Scientific domains at large.



(b) Turning points in our understanding.

Figure III.5.5: *The structural hierarchy unravelled by the sciences.* On the left in the circle it is basically the gravitational force that causes structure, while on the right it is due to the other forces. Top down we see basically how time evolution both from the left (from large scales down) and the right (from small scales up) lead to ever more complex structures. The turning points in our understanding of nature can also be mapped on the circle.

awareness of the social impact they may have. It is our duty to educate a critical audience, that is conversant about topics that will shape our common future. In my opinion those topics include the possible ways in which we may steer and regulate future applications of science and technology so that they improve the human condition not for the few but for the many.

What adds to the complexity of this process is the fact that the plusses and minuses of novel technologies are in most cases not evident at the moment of their inception. Unfortunately they are often even intertwined. And that is precisely why the incorporation of up-to-date scientific expertise in the political arena is necessary in any well-functioning, future oriented democracy.

Trees of knowledge

What we learned in this process of scientific discovery is presented schematically in a series of four subsequent images. You may call it a display of the harvest of the double helix.

The structural hierarchy. In the first picture III.5.4(a) we mapped the structural hierarchy of the material world onto a circle, where moving clockwise we go to ever larger distances. At the bottom, roughly in the middle, we see ourselves, and it is from that position that we started to explore the order of things in- and outside of us, diving ever deeper in the microcosmos and looking ever further out in the macrocosmos. So one way to look at this figure is that it depicts the human effort to understand the world we are living in, basically following the double helix of science and technology.

Three fundamental frontiers. The arrows we superposed on the circle in the second picture III.5.4(b) indicate how the basic frontiers of knowledge have moved forward. On the left from starting with Galileo all the way up to the Hubble or Webb space telescopes, and on the right from Antonie van Leeuwenhoek all the way down to the LHC at CERN. Very large and very small scales meet and merge in the Big Bang where modern research fields like astroparticle physics came to flourish. The Big Bang is the event where today's largest and smallest scales of the universe meet and that is why I have put the scales on a circle and not on a line.

The inside arrows pointing down to us humans clearly represent the evolutionary perspective on structural complexity like the phenomena of life. The arrow on the left represents the study of biology from the macroscopic Darwinian perspective on the speciation of plants and animals, and on the story told by the fossils they left behind in the earth's crust. The downward arrow on the right represents the unstoppable advance of molecular thinking in the life sciences, symbolized by the DNA-molecule. And indeed the genes on the DNA molecules tell that same Darwinian story but then on the molecular level. These two complementary views on evolution therefore meet and merge in the modern life and the earth sciences. And in a sense this 'closes' the circle at the bottom in us humans.

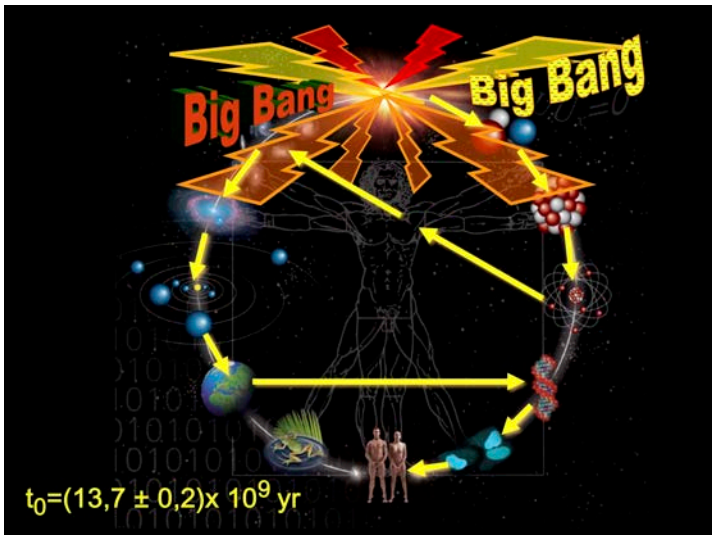
Three domains: Relativity, Quantum and Evolution. As indicated in the third picture III.5.5(a), the arrows in the background represent the large domains of fundamental scientific inquiry which are anchored in the leading conceptual frameworks like the *domain of relativity* (concerning space-time and gravity), *the domain of quantum* (covering all forms of constituent matter and the forces between them), and finally the *domain of evolution*, the concerted effort to gain a unified understanding of the tremendous diversity and complexity that evolved in nature over time.

Quantum versus Relativity. Quantum theory is less accessible than relativity, because as we saw it is the impressive legacy of a great number of outstanding scientists that filled over a century of successful groundbreaking research. For that reason quantum has not been personalized to the degree that relativity has been identified with the person of Albert Einstein, and maybe that also explains why intellectual giants like Bohr, Schrödinger, Heisenberg and Dirac never reached the status of a public idol like Einstein. The painful paradox is that whereas their profound work is leaving ever deeper marks in modern life, most people bitterly complain that they do not understand a single word of it. And that was one more reason to write these books.

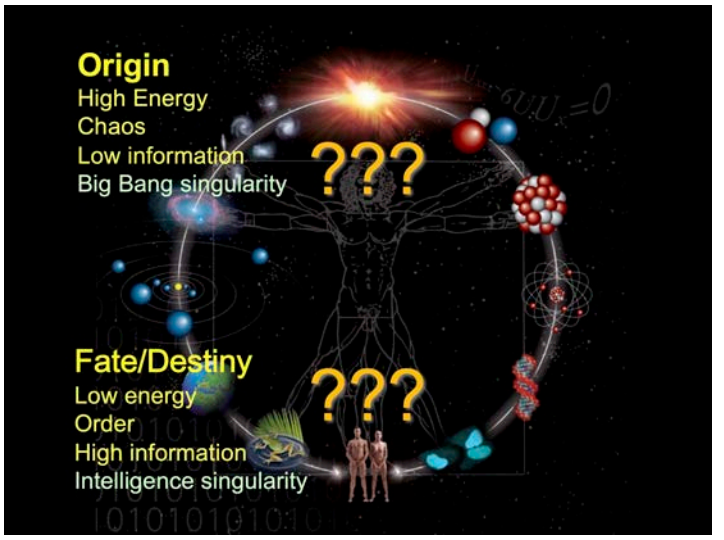
It is interesting to note that a Nobel prize for the theory of relativity as such has never been awarded, while there have been more than fifty linked to quantum theory as witnessed by the tables in appendix B on 'Chronologies, ideas and people.' Indeed, the prize awarded to Einstein, was in recognition of his explanation of the photo-electric effect, which is a fundamental contribution to quantum theory and has nothing to do with relativity. So the irony is that he received the Nobel prize for his contribution to a theory he basically didn't believe in!

With so many Nobel prizes awarded, it is no surprise that a book that aims slightly higher than just summing up the basic results is bound to be voluminous indeed. Be my guest!

Turning points. In the fourth Figure III.5.5(b) we show how this endeavor to advance knowledge gave rise to a rather limited number of truly fundamental turning points that stand for the great leaps forward in our scientific understanding of the natural world, the world we ourselves are part of. It is striking to see that there are only so few. It is also striking that so much novel science and technology derives from such a small number of truly fundamental insights.



(a) The Big Bang and the subsequent cosmic evolution.



(b) Ultimate questions that concern our deep origins and our long-term future (if we have one).

Figure III.5.6: *Science appears caught between two singularities.* The cosmic evolution at large according to the Big Bang scenario is depicted in Figure (a). The ultimate questions in Figure (b) concern on the one hand the origin somehow hidden in the Big Bang, and on the other is about where this evolution will bring us and to what extent we can shape that future ourselves. So it concerns nothing less than the quest for the interpretation or meaning of our universe as a whole, and its present and possible future contents.

Cosmic evolution. Let us continue with the two pictures of Figure III.5.6. In the first one we depict the actual process of cosmic evolution according to the hot Big Bang scenario. Where the increasing complexity in dead matter smoothly turns into the Darwinian story of life. This took altogether almost 14 billion years, where the Darwinian episode 'only' covers the last 4.5 billion years. Clearly the full story is by no means complete. The figure nicely shows how material complexity sequentially evolved as a necessary consequence of an expanding universe slowly cooling down. It is this story of cosmic evolution that brought most of the empirical natural sciences together so harmoniously, that makes the narrative or perspective of science on the whole of nature so clarifying and illuminating. It is in that story that reductionism meets holism. A beautiful product of brainpower, enlightenment and perseverance.

Ultimate questions: from origin to fate. Science is a systematic process of advancing understanding by creating ever better observational abilities, which in turn allow for ever better modelling of reality. The circle that appears in all the figures by no means tries to convey the idea that science is a closed body of knowledge, a narrative completed. Science is always 'work in progress,' and may on the one hand be characterized by the questions it *did* answer, but on the other hand by the questions it raised but did *not* answer. This is indicated by the question marks at the top and bottom of the would-be circle. They represent ultimate questions that in fact rip open the circle allowing for additional realities we have not yet any idea about. It illustrates how the whole of science is basically caught in between two essential but enigmatic singularities.

On top we have what I called the 'cosmic short' between

the physics of the smallest and largest conceivable scales which somehow meet in the Big Bang. We like to think of the Big Bang as an event, but may be it is better to think of it as a gate to an unknown territory where relativity and quantum presumably govern in a truly unified fashion. In that point there is room for fundamentally new insights. That gate would give access to the physical origins of the Big Bang itself. Our lack of understanding is probably best characterized by the term 'Big Bang singularity,' which of course refers to the unphysical extrapolation of the early universe to the quite unphysical initial state with an infinite temperature and energy density.

The arrows of time move downward towards the domain of human evolution, of the human brain, and of human society. Clearly also at that point our understanding is very much incomplete. The present state of science poses hard questions, like asking how the process of evolution will further unfold. It is a fact that the theory of evolution, in spite of having an incredible explanatory power with respect to our past, is surprisingly weak as a predictive model. It predicts a process of the increasing complexity of organisms but is not specific about where the breakthroughs of – let us call it – biological self-transcendence will take place. And this question of predictability has not become easier as we humans have become the dominant species on Earth. As indicated in the figure we have moved from an initial state, which is characterized by extremely high energy, chaos, a uniform distribution of a low information content or capacity, towards the present state which has the signature of very low temperature and energy, allowing for highly localized forms of complex order and high information capacity like the brains of human beings for example.

Evolution at large. In Figure III.5.7 I have presented an alternative visualization of the cosmic evolution at large and marked the most consequential branchings of the evolutionary tree. I like to think of these branchings as moments of radical innovation, as irreversible transitions or tipping points. Indeed, we went through the evolution of dead matter all the way up to the production of the chemical elements which were a necessary prerequisite for the creation of sustainable life on Earth and may be elsewhere on what are called *exoplanets*. In a universe with some 10^{21} stars that probability of extraterrestial life can't be negligible I would think.

To cope with the unknowns of the future a solid knowledge of our past appears to be a crucial prerequisite. So, we should celebrate collaboration in scientific research efforts addressing such questions, like the launch and operation of the James Webb space telescope that allows us to look deeper in the universe than we ever did before, exactly to better understand its remote past. It is a splendid international collaboration of NASA and the European and Canadian Space Agencies. Its mission is to collect hard data concerning the beginning of structure formation and the births of stars as well as the possibility of extra terrestrial life (see Figures III.5.8 and III.5.9).

Once life began, we had another 4.5 billion years of biological evolution culminating in such attributes as consciousness and intelligence which allowed humanity to basically take over their planet. Human evolution transformed us from just inhabitants to the custodians of planet earth. It appears that we have taken our fate in our own hands. We have become responsible for our own future. At present that means that we have to face such inconvenient truths like the climate crisis, and we need to urgently act in order to keep the planet inhabitable. Al Gore, the former vice president of the US and a powerful voice in favor of direct action to avoid climate catastrophes, once noted that by broadcasting an inconvenient truth one is bound to wake up the most powerful enemies, which makes taking proper action even harder.

We also must seriously analyse the consequences of the great information revolutions that obey Moore's law, and the introduction of internet and its radically novel way of

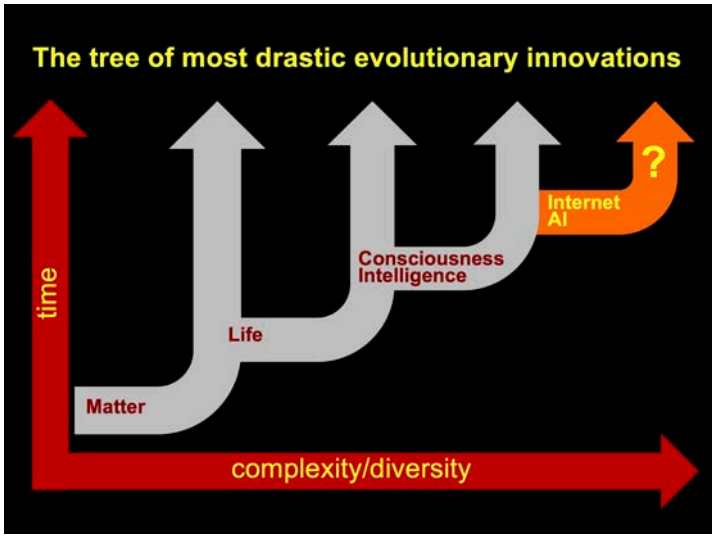


Figure III.5.7: *Cosmic evolution at large.* Does human evolution, driven by the double helix of science and technology, allow for a post-biological branch dominated by artificial intelligence, machine learning and quantum computing?

'connecting people.' The introduction of these new technologies that allow for instantaneous and global human interaction clearly implies a fundamental change in the human condition which has caused a tipping point in social awareness and coherence. It started a process of a global restratification of society, and the unfolding of unheard of concentrations of power and wealth. This process is full of social risks and has to be critically monitored and controlled by governments and international institutions that should be endowed with both sufficient funding and executive power. This is a far cry from today's reality.

To cope with the many negative aspects of these developments requires the development of the notion of global citizenship. People should be educated to be aware of what is happening, and institutions should insist on openness, accessibility and transparency. This may necessitate adding new chapters to the declarations of fundamental human rights, which extend and define these rights to their existence on the World Wide Web and other cyberworlds. It teaches us, as the dominant inhabitants of planet Earth, that the tremendous amount of freedom we have achieved implies a huge undeniable responsibility.

A post-biological branch? Information philosophers and futurists like Max Tegmark, Nick Bostrom and Yuval Harari warn us that with the rapid advances in artificial intelligence, like machine learning, and quantum computing, machines may well take over completely as we become more and more dependent on them. Not just for gathering relevant information, but also for making rational, optimal decisions. There are major obstacles to be taken, namely, to extend the abilities of artificial intelligent algorithms to have 'general intelligence.' This is a much harder problem than acquiring expertise in a limited context and domain in which algorithms already outperform humans. General intelligence is the outcome of our biological evolution and unsurprisingly, that is what humans excel in.

Anyway, the question posed by the orange branch in the figure is whether we are on the verge of a transition towards a radically different post-human, post-biological evolutionary phase. This does not mean that we could no longer exist, bacteria after all managed to survive in many ways too well for billions of years after more complex organisms took over. What the post-human branch presumably implies is that we are no longer the glamour boys of creation, but rather that we may turn into somewhat outdated pieces of biological apparatus of reduced relevance, compared to our super intelligent silicon or quantum brothers and sisters to be. Maybe the optimal way forward is to engage in further exploring symbiotic options.

The intrinsic value of science. We should be aware that politicizing science is a threat to its primary objective: the search for objective truths. The risk of trying to politicize that aspect is not just that it leads to crimes against logic, but also to corrupting scientific integrity. It often involves a form of 'passive lying,' which refers not to directly telling plain lies (active lying) but rather to not telling the truth, that

is, the whole truth. It is like leaving important terms out of the equation and thus propagating models that fail reality. It is like the often-applied strategy of spreading misinformation to gain political or commercial support and influence. 'The goal justifies the means,' is the slogan that easily comes along and allows the most well-funded lobbyists to dominate the political landscape. Indeed, the success of advertising is justifying the goal of better sales often by not telling the truth.

But doesn't science do the same, you might object? Yes and no! It is certainly true as I have noted repeatedly in the book that science is 'work in progress,' and therefore also scientific 'truths' are relative and should be subject to refutation if decisive arguments or data are being brought forward at some point. Indeed, the notion of an absolute truth is basically incompatible with the notion science as an incomplete body of knowledge. And it is this aspect that makes the scientific infrastructure, its institutions and funding strategies vulnerable to abuse. This is a paradoxical aspect of the role that science plays in society: although there is no such thing as an absolute truth, we do not hesitate to board planes, go to hospitals, and get addicted to our cell phones. It appears that scientific truths, if not absolute, are at least extremely robust!

The symbiotic relationship between science and technology is harder to disentangle. As stated before they need each other in essential ways, and yet technology is per definition a double-sided sword. The best we can do is to insist that the discourse on science and technology at all stages be a hundred percent transparent and respects the principles of a solid democracy. This refers to a higher vocation, and adds elements of ideology and wishful thinking to the notions of science, technology and innovation, which in turn make them more vulnerable!

In my opinion what we need is quite the opposite of what is trending: we need to have more science, scientific literacy and expertise into the political arena to bring the necessary amount of integrity into the political discourse. Unfortunately science as the evidence-based cornerstone of human culture remains a vulnerable institution that should be protected and defended against the arrogance of power, media popularity, the spreading of misinformation, and lobbying practices that turn into corruption. In the words of the well-known spy novel author John le Carré:

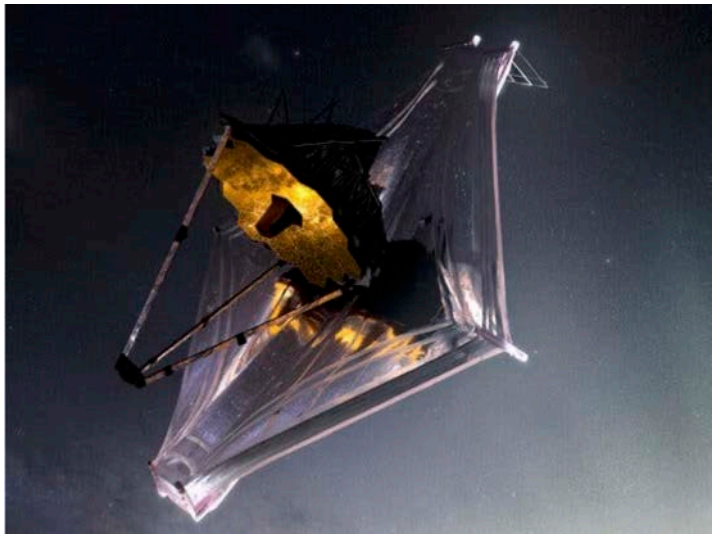


Figure III.5.8: An artist impression of the James Webb space telescope (JWST) unfolding in space at 600.000 km from the Earth. Its mission is to look at the very early stages of the universe as a whole and the very early stages of structure formation. It is furthermore the first space telescope to study the possibility of extraterrestrial life by analyzing the chemical composition of the atmosphere of exoplanets. The slogan would be: Are there somewhere in the universe alternative humankinds? (Source: Adriana Manrique Gutierrez/NASA)

> One day somebody will explain to me why it is that, at a time when science has never been wiser, or the truth more stark, or human knowledge more available, populists and liars are in such pressing demand.
>
> *John le Carré*

Indeed, as soon as we allow the politicization of the fund-

ing structure, and make it a prey to lobbyists and commercial interests, or force it to serve the vested interests and privileges of some ruling class, irrespective of the political system we adhere to, we are sure to lose science. It will decay from a devoted search for truth – also if that truth turns out to be inconvenient – to some kind of hidden or even blatant form of lopsided advocacy. Legal experts or lawyers, in contrast to scientists, are allowed to limit their sources for research and part of their skill requires craftily selecting the evidence that supports their client's case.

Here is a large-scale perspective offered by the eminent quantum scientist Charles Bennett:

> The Enlightenment inspired Universal Declaration of Human Rights promulgated in 1948 after a decade of technical sophistication accompanied by inequity and cruelty on an unprecedented scale, exemplifies the seemingly still attainable goal of an equitable, peaceful society that manages its environment and itself well enough to last millions of years.
>
> *Charles H. Bennett*

Human history looks like a perpetual battle between power and knowledge, with power always calling victory in the short term (under the argument of improving efficiency and 'the' economy) and knowledge always being the winner in the long term, even though the price for society for finding out can be disproportionally high. We created dangerously pervasive constructs like the military-industrial complex, or the medical-industrial complex and now also the information-industrial complex, which have turned into autonomous self-inflating entities thoroughly intertwined with human society. These thrive on a delicate interplay between innovation and commercialism using the creation of fake needs and fake fear. They embody an abuse of power that is derived from knowledge. The sobering fact is that lies and misleading accounts spread fast and one can only hope that truth will ultimately prevail. I myself firmly believe that to be the case, but overall it remains an open question. Too much science/technology-based power in too few hands is a recipe for societal disasters. Let me close with quoting Bennett once more:

> Unfortunately, due largely to the increased range and speed of communication, misinformation has emerged as a meta-threat to equity and civilisation. By luring people into self-isolating bubbles, to be soothed, entertained and incited by incompatible versions of reality, it empowers autocrats and demagogues, it hobbles democracies and makes cooperation on globally urgent problems like climate change almost impossible.
>
> *Charles H. Bennett*

Addressing scientific illiteracy.

> *Heisenberg? Huh, isn't that the guy from* Breaking Bad*?*

After the red light started flashing, the radio host nodded to me and asked: 'Well, professor, can you tell us in a few lines what quantum physics is?' And I said: ' Hm, yes of course, hmm I mean No! Hmmm, I mean yes, but ...' Talking quantum to family and friends at a birthday party often feels like being a tour guide in London for extra-terrestrials who don't happen to know what a bridge, a museum or a traffic light is. As I mentioned before, the fact that quantum things are largely invisible does not mean that they are not there. They certainly are. And as we have learned, the fact that most quantum things are not discernible by the naked eye doesn't mean that they are not relevant or important. In spite of being unknown and widely ignored, the *quantessentials* are here to stay. This leaves us with the sobering fact that they are still surprisingly unfamiliar. This in my opinion is a strong call for worldwide efforts to educate, to fully develop the tremendous intellectual potential that is present everywhere at any instant.

Figure III.5.9: *Starbirths in the Carina Nebula as seen by the James Webb space telescope.* This image made in July 2022 is divided horizontally by an undulating line between a cloudscape forming a nebula along the bottom portion and a comparatively clear upper portion. Speckled across both portions is a starfield, showing innumerable stars of many sizes. The smallest of these are small, distant, and faint points of light. The largest of these appear larger, closer, brighter, and more fully resolved. The upper portion of the image is blueish, and has wispy translucent cloud-like streaks rising from the nebula below. The cloud-like structure of the nebula contains ridges, peaks, and valleys - an appearance very similar to a mountain range. (Source: NASA, ESA, CSA, and STScI.)

I have spent about half a century in that invisible quantum world, doing a lot of active research, but also getting slightly frustrated not being able to share much of it at everyday occasions like birthday parties. At times that made me sad but also aware that I should stop whining and just sit down and write a book about what I learned on my journeys through that amazing quantum world. A modest attempt to help alleviate the burden of scientific illiteracy. And that is how the three lines allowed to me by that sympathetic interviewer gave rise to these three volumes about the *Power of the Invisible: The Quantessence of Reality*.

Further reading.
Classics of popular physics:

- *Cosmos*
  Carl Sagan
  Random House (1980)
- *A Brief History of Time*
  Stephen Hawking
  Bantam Dell Publishing Group (1988)
- *Cosmic Code*
  Heins Pagels
  Dover Publications (2012)

On Science and the Future of Human Culture:

- *Superintelligence: Paths, Dangers, Strategies*
  Nick Bostrom
  Oxford University Press (2016)
- *Sapiens*
  Yuval Harari
  Penguin books (2015)

Complementary reading:

- *A Project to Find the Fundamental Theory of Physics*
  Stephen Wolfram
  Wolfram Media (2020)
- *In Praise of Science: Curiosity, Understanding, and Progress*
  Sander Bais
  MIT Press (2010)
- *Mysteries Of The Quantum Universe*
  Thibault Damour and Mathieu Burniat
  Penguin (2020)

Appendix A

Math Excursions

♣ On functions, derivatives and integrals

> Do not worry about your difficulties in mathematics.
> I can assure you that mine are greater still.
>
> *Albert Einstein*

Functions. Functions are a general class of objects in mathematics that have endless applications in all fields of science. A function is an object – let us denote it by the symbol f – that may depend on a set of variables (arguments) – say $\{x_a\}$. As such it assigns a value to f for any allowed point in the space of variables $\mathcal{X} \sim \{x_a\}$: in other words it provides us with a map $f : \mathcal{X} \to \mathcal{F}$. The domain $\mathcal{F}$ of the function denotes the space where f itself lives, and can be many things, we think in particular of the real numbers $\mathbb{R}$, the complex numbers $\mathbb{C}$, or some (other) vector space $\mathcal{V}$.[1]

Think of the temperature T in the room you are in. It is a function that depends on where and when, i.e. on the set of variables $\mathcal{X} \sim \{x, t\}$, you could say $T : \{x, t\} \to \mathbb{R}$ and we indicate this dependence by writing $T = T(x, t)$. The potential energy $V(x)$ of a particle is a real function defined over the real position space, and like the temperature, V may differ from place to place. If we plot the value of a real function f as the 'height' above the point x then $f(x)$ defines a kind of *landscape* over $\mathcal{X}$. Very basic features of functions are given in Figure A.1 which refer to whether they are continuous and or differentiable. We will

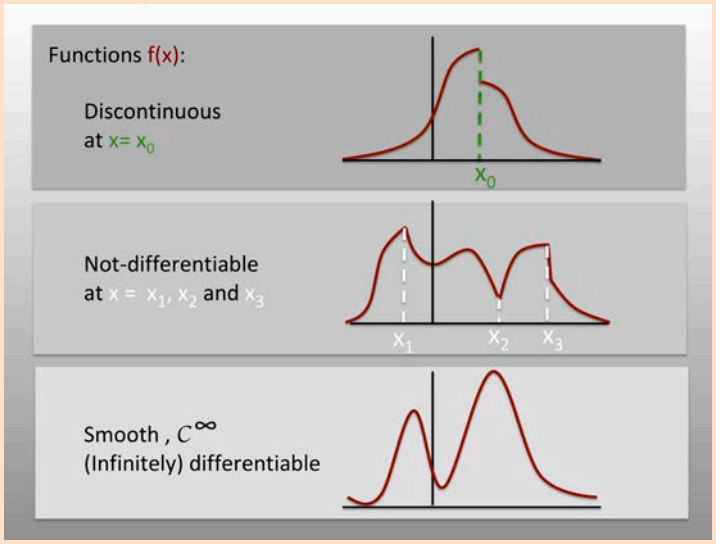


Figure A.1: *Function classes.* We have plotted three functions which belong to different classes. A *discontinuous* function on top (the function value jumps at $x = x_0$). In the middle a continuous function but *not-differentiable* at $x = x_1, x_2$ and $x = x_3$, where the slope is discontinuous when approaching the points from the left and the right. At the bottom a *smooth* function which is per definition *infinitely differentiable*, meaning that all higher derivatives exist and are continuous.

[1] We mention the words 'complex numbers' and 'vectors' here just in passing; these notions are discussed in later Math Excursions.

mostly assume that we are dealing with *smooth* functions: those are functions for which all derivatives exist and are continuous.

We have mentioned other quantities which are basically functions: the position and velocity are functions of the time variable. In d-dimensions these are vectors ("vector" functions) with d components. Vectors have not only a magnitude but also a direction which makes them different from being just a number. A number can be written down and be communicated by mail; this is not true for a vector because the direction can get messed up. The electric and magnetic fields are both vector-valued functions or vector fields in short. The same is true for the velocity field of a river, it encodes the direction in which the fluid flows at any given point in the fluid. So even if you were not aware of the notion of (vector) functions, you presumably now realize that you are quite familiar with them. To give you an impression, we have plotted some typical elementary (real) functions of a single variable in Figure A.2.

With real functions you can do what you can do with numbers if you do it point wise, i.e. in every point of $\mathcal{X}$. For example, we define the product h of functions f and g by the function $h(x) = f(x)g(x)$. The limitations on what you do with functions is of course determined by which operations are defined in $\mathcal{F}$.

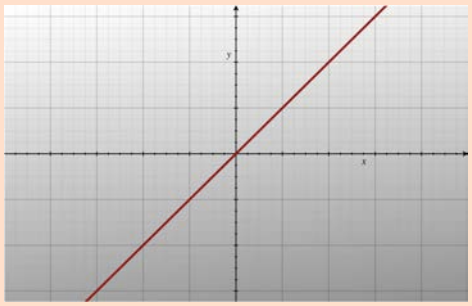
Of interest are two natural operations one may define on smooth functions that play a fundamental role in many applications. These operations are basically each other's 'inverse'; one is called *differentiation* or taking a derivative, the other is *integration*, or taking the integral. We discuss them for the case of real functions.

Differentiation. Think of a real function $f(x)$ of one real variable, then we may draw it as a curve on a graph paper, putting x along the x-axis and $f(x)$ along the y-axis, as we did in Figure A.3(a). The derivative with respect to the variable x in a point x_0 of the function denoted as $\frac{df}{dx}$, or simply with a prime, i.e. $f'(x_0)$ is just the *slope* of that curve above the point x_0.

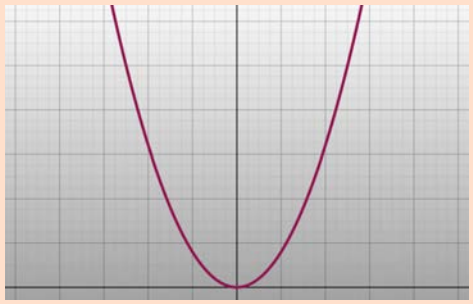
For example, if the function is linear in x , $f(x) = 3x$, then that function has a constant slope equal to 3 and thus is the derivative a constant, $f'(x) = 3$. Having given this heuristic definition of the derivative, I should hasten to say that this is a phenomenally important concept in science, as it embodies the mathematical statement that exactly quantifies the otherwise rather vague notion of 'change'.

Looking at the derivative operator more abstractly it can be considered as a map $\frac{d}{dx} : \mathcal{F} \to \text{Slope } \mathcal{F}$. Points where the derivative of a function vanishes correspond to points where the slope is zero and the function has a maximum or a minimum, as we have indicated in Figure A.3(a). Note that if one knows a function in the neighborhood of a point x_0 one may calculate its derivative in that point. This is clear from the formal definition of the derivative: $f'(x) = (f(x + \Delta x) - f(x))/\Delta x$ taken in the limit of ever smaller Δx. This definition implies another useful relation (also in the small Δx limit) namely that we may write: $f(x + \Delta x) = f(x) + f'(x)\Delta x$. This provides a clear statement of the use and meaning of a derivative: if we make a tiny move from x to $x + \Delta x$ in space, then the corresponding change in any function $f(x)$, is from $f(x)$ to $f(x) + f'(x)\Delta x$ to lowest order in Δx.

Let us finally mention that calculating the derivatives of many standard functions and expressions containing them is not so hard and usually part of a science high school math curriculum. We have listed a few derivatives of standard functions in Table A.1 below. Another way to think about differentiation is therefore to say that it is an operator d/dx which applied to a function $f(x)$ generates a translation (or change) in function space $\mathcal{F}$ induced by a small translation in the underlying configuration space $\mathcal{X}$. We will make use of this interpretation later on.

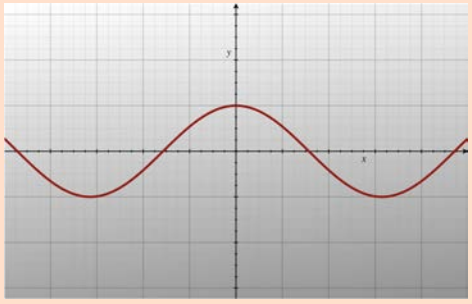


(a) The linear function $f(x) = x$. It has a constant slope. It is the simplest *odd* function as it satisfies $f(-x) = -f(x)$

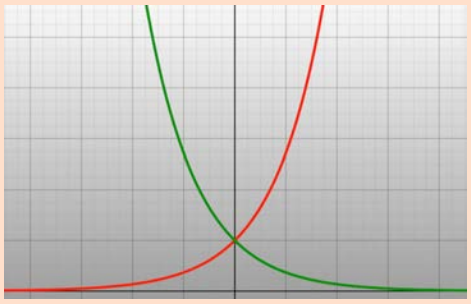


(b) The quadratic function $f(x) = x^2$. It has a constant curvature or second derivative. It is the simplest even function (except the constant function) satisfying $f(-x) = f(x)$.

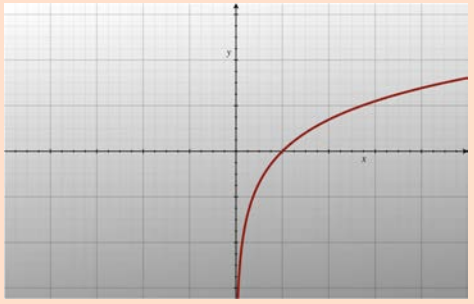
(c) The inverse function $f(x) = 1/x$. The function slowly tends to zero for $x \to \pm\infty$, while it becomes infinite (or singular) for $x \to \pm 0$. It is only defined for $x \neq 0$.



(d) The periodic function $f(x) = \cos(x)$. It satisfies the property $f(x) = f(x+2\pi)$. Shifting the cosine by $1/4$ period to the right one obtains the sine function.



(e) The exponential functions $f(x) = e^{\pm x}$. these grow rapidly to ∞ for $x \to \pm\infty$ and decay rapidly to zero for $x \to \mp\infty$.

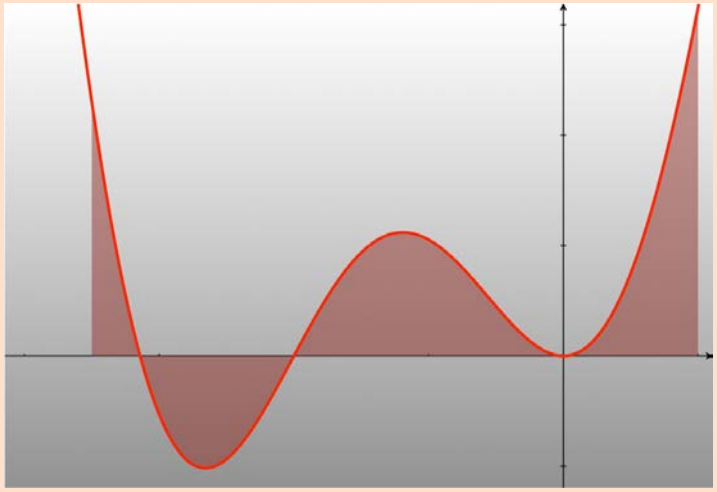


(f) The logarithmic function $f(x) = \ln(x)$ is a slowly but ever-growing function. It has a singularity for $x \to +0$.

Figure A.2: The graphs for some typical elementary real functions f(x), showing their salient features.

An example: dispersion. We have been discussing the energy E of a particle as a function of the momentum p for the non-relativistic and relativistic cases with a parametric dependence on the mass m_0. There is another quantity of importance and that is the *dispersion* defined as the derivative of E with respect to p. The term dispersion originates in optics where in a given medium one has that the frequency will depend on the wavelength, which manifests itself for example in the fact that the angle of refraction of light will depend on the angle of the incident beam. For matter waves we have that $E = \hbar\omega$ and $p = \hbar k$, so we can express the dispersion also in terms of E and p. In Figure A.4, I have plotted the relativistic expression for the particle energy, $E = \sqrt{m_0^2c^4 + p^2c^2}$, and below it the dispersion $dE/dp = pc/\sqrt{m_0^2c^2 + p^2}$. There are roughly three regimes: (i) on the left we have the non--relativistic regime where $p \ll m_0c$ where the energy approximates tp $E \simeq m_0c^2 + p^2/2m_0$ with linear dispersion $dE/dp \simeq p/m_0$, and the expression up to the mass-energy reduces to the familiar Newtonian form, (ii) in the

(a) The *derivative* df/dx (purple) of a function $f(x)$ (red). At the extrema of $y(x)$ the derivative (= slope) is zero.



(b) The *integral* $\int_a^b f(x)\,dx$ of $f(x)$ is the area below $y(x)$ above the x-axis minus the area belowthe x-axis, between the points $x = a$ and $x = b$.

Figure A.3: A function, its derivative, and its integral.

middle we need the fully relativistic expression, and (iii) on the right we have the ultra-relativistic regime where $p \gg m_0 c$, and we have the approximation $E \simeq pc$ with dispersion $dE/dp \simeq c = constant$, which effectively corresponds to the expression for a massless particle.

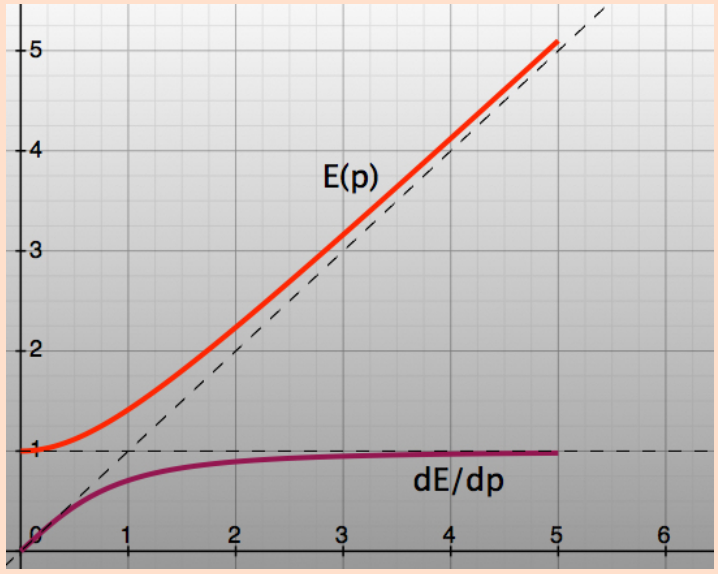


Figure A.4: *Relativistic energy.* The particle energy E as a function of p in red, and the dispersion defined as the derivative dE/dp in purple. We have chosen m_0 and c equal one.

Integration. Having the red curve in the example of Figure A.3(b) the (definite) integral F_{ab} of a function $f(x)$ between two points $x = a$ and $x = b$ is just the area under the curve between the two points. One may also define an 'indefinite' integral $F(x)$ or primitive of $f(x)$, which is mathematically represented by the integral symbol:

$$F(x) = \int f(x)dx\,. \tag{A.1}$$

F(x) has the property that $F_{ab} = F(b) - F(a)$. If a function is constant $f(x) = c$ then the integral is thus simply $F_{ab} = c(b - a)$ and $F(x)$ would be $F(x) = cx + d$ where there is an arbitrary constant d that one can add. Now we are also in a position to appreciate the remark that these operations are in a sense each other's inverse: if we differentiate $F(x)$ we get the original function $f(x)$ back.

The definition of the integral involves a limiting procedure of an approximation that is not so hard to imagine. To calculate the definite integral F_{ab}, we divide up the interval

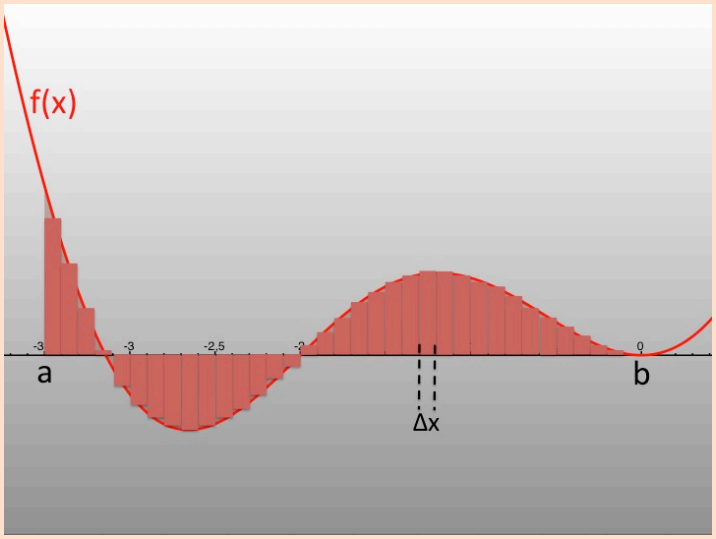


Figure A.5: *The integral as area.* The definition of the definite integral of $F(x)$ between points $x = a$ and b is the sum of the positive and negative contributions from the areas of the small rectangles, in the limit that $\Delta x \to 0$.

$b - a$ on the x-axis up in a large number N equal little segments Δx, then we define the centre of each segment by its coordinate $x_i : i = 1, \ldots, N$. The integral is then defined by:

$$F_{ab} = \int_a^b f(x)dx = \lim_{N\to\infty} \sum_i^N f(x_i)\Delta x\,, \qquad (A.2)$$

as is illustrated in Figure A.5.

Calculating the integrals of elementary functions is not too hard, but often integrating is hard and not possible in 'closed form'. Therefore, numerical approximations are of crucial importance in most applications, and those are usually based on approximations in the spirit of equation A.2. The problem of integration is at the heart of physics and engineering, exactly because in most cases the laws that govern nature are formulated as so-called differential equations, that means that the equations contain derivatives of quantities one would like to solve for. Many equations are 'equations of motion'. The equations of Newton determine the time evolution of a particle's position end momentum. Those of Maxwell do that for the electromagnetic fields, and the Schrödinger equation for the wavefunction of a quantum system, while Einstein's equations describe the time evolution of the universe. Solving those equations corresponds in some sense to finding ways to 'integrate' the equations for specific boundary or initial conditions.

derivative : $\frac{df(x)}{dx}$	**function :** $f(x)$	**integral :** $F(x) = \int f(x)dx$
a	ax	$\frac{1}{2}ax^2$
nx^{n-1}	$x^n (n \neq -1)$	$\frac{1}{n+1}x^{n+1}$
$\frac{-1}{x^2}$	$\frac{1}{x}$	$\ln\lvert x\rvert$
$\cos(x)$	$\sin(x)$	$-\cos(x)$
$-\sin(x)$	$\cos(x)$	$\sin(x)$
ke^{kx}	e^{kx}	$\frac{1}{k}e^{kx}$
$\frac{1}{x}$	$\ln x$	$x\ln x - x$
$\Longrightarrow$ differentiation	$\Longrightarrow$ $\Longleftarrow$	Integration $\Longleftarrow$

Table A.1: A list of some elementary functions (see also Figure A.2) in the middle column, with their derivatives on the left and their integrals or primitives on the right. Taking a derivative moves you to the left, integrating moves you to the right. Integration means that one always can add an arbitrary constant to the integral; this constant is not included in the table.

In Table A.1 we have listed some well-known functions,

their derivatives, and their primitives (i.e. integrals).

Example: the harmonic oscillator. The force on a particle is defined as minus the derivative of the potential energy: $F = -dV/dx$, indeed with $V = ax^2/2$, this yields the harmonic force $F = -ax$. But given the force we can also calculate the potential energy by integrating it. We have to move the particle up the hill form $x = 0$ to, say, $x = x_0$. To do so we must do an amount of work on the particle which equals the (opposite) force times the distance, integrated from zero to x_0:

$$V = -\int_0^{x_0} F(x)\, dx = \int_0^{x_0} ax\, dx = \{\frac{1}{2}ax^2\}_0^{x_0} = \frac{1}{2}ax_0^2 .$$

Differential equations. Differential calculus is basically the calculus of changes, and differential equations are typically the equations that govern the change in time or space of any dynamical system one might think of, equally applicable to modelling in classical physics as it is for quantum theory, but it is equally well employed in modelling economics, ecological systems or the climate. As we have seen, many 'laws of nature' take the form of a system of differential equations. This means that on the left-hand side of the equation we have the changes of the system's variables in time and space, while on the right-hand side they are expressed as functions of the variables themselves, i.e. the point in the space of states the system could be in. Examples were already provided by Newton's equations (I.1.3) and the Maxwell equations (I.1.28). The solutions of these equations describe therefore the dynamical trajectories in the configuration space that the system traverses in time. The trajectory depends of course on the starting point or initial condition. Obtaining solutions to differential equations has to involve some kind of integration because we want to get rid of the derivatives, and that is exactly what makes solving differential equations so hard. If the equations are linear, meaning that the unknowns one want to solve for only appear linearly in the equation, solutions can often be obtained in closed analytic form, but if the equations are nonlinear that is only rarely the case.

Let us conclude this excursion by looking at two differential equations of particular interest, a growth/decay equation and a wave equation.

Example: the equation for exponential growth or decay. We have a container with N_0 radioactive nuclei. Then the remaining number $N(t)$ at time t will decrease in time at a rate dN/dt. This rate will be proportional the number N(t), which is just saying something like, 'if the population is twice as big, twice as many people will die.' So the equation we like to solve reads:

$$\frac{dN}{dt} = -\lambda N . \tag{A.3}$$

this can be cast in the form:

$$\frac{dN}{N} = -\lambda dt . \tag{A.4}$$

Now the left-hand side and the right-hand side can be 'integrated', which by using Table A.1 yields the solution:

$$\ln|N| + d = -\lambda t \;\Rightarrow\; N(t) = N_0 e^{-\lambda t} , \tag{A.5}$$

where the constant e^{-d} has to equal N_0, the number of nuclei at time $t = 0$. Note that the solution corresponds to the green curve depicted in Figure A.2(e). The solution tells us that the decay is exponential, and we will refer to this result if we talk about radio-active decay in chapter I.4. And if we change the sign in front of λ in the equation, we of course get the red curve in the figure corresponding to exponential growth, describing some stages of epidemics or a post on Facebook 'going viral.'

Example: the wave equation. This equation is of interest because waves appear all over the place in physics. Not just water or sound waves, also light is a wave phenomena, and also in quantum theory we encounter wave equations in many guises. Most prominent is the Schrödinger equation, but also the Maxwell and Dirac equations

are basically wave equations, which after quantization will have interpretations in terms particles. And it is here that the well-known quantessential catch phrase *particle-wave duality* originates.

In one space and one time dimension the relativistic wave equation takes the form of a differential equation with two derivatives working subsequently on a function $f(x,t)$ of space and time:[2]

$$\frac{\partial^2 f}{\partial t^2} - c^2 \frac{\partial^2 f}{\partial x^2} = 0\,. \tag{A.6}$$

The solutions for f are waves that move with a velocity equal $\pm c$ for example:

$$f(x,t) = a\cos(\omega t - kx)\,. \tag{A.7}$$

This solution has besides the *amplitude* a, two parameters, the *angular frequency* $\omega = 2\pi\nu$, and *wavenumber* $k = 2\pi/\lambda$, and looks like the wave pattern of Figure A.2(d) moving either to the left or the right. Indeed, taking two derivatives means in Table A.1, that we move from the column on the right to the column on the left. If you put this into the equation and take the derivatives, you get an algebraic equation $\omega^2 - c^2k^2 = 0$ for the parameters ω and k, telling us exactly, that – as advertised – there are propagating waves satisfying the equation with $\omega = \pm ck$ which amounts exactly to the wave relation $\nu = c/\lambda$. Later on we will see that quantization of this relation leads to the linear dispersion relation $E(p) = \hbar\omega = c\,\hbar k = cp$, which is characteristic for a massless particle. This reflects the similarity of the above equation with the electromagnetic wave equation (I.1.47). ♣

[2] As f depends on two variables we have to distinguish the derivatives with respect to space and time, we write the curly derivative symbols called *partial derivatives*. The squares in the derivatives mean that you apply the derivative operator twice, so $\partial^2 f/\partial t^2 = (\partial/\partial t)^2\, f$.

♢ On algebras

In high school we have to learn *elementary algebra*, where one represents variables – mostly corresponding to real numbers – as abstract letter symbols, and one learns how to manipulate the expressions according to certain rules or operations that apply to real numbers, such as addition and multiplication. The principal application is to solve equations by exploiting these manipulations. For example, having the quadratic equation $ax^2 + bx + c = 0$, the question is to solve for the variable x in terms of the constants a, b and c. One proves that there are two real solutions given by $x_\pm = (-b \pm \sqrt{b^2 - 4ac})/2a$, provided the expression under the square root is positive. So the advantage of the abstract notation is that the answer applies for any choice of the constants a, b and c: it gives the general solution.

Abstract algebra. Generally, the subject of *abstract algebra* deals with collections of objects such as numbers, vectors, matrices, polynomials and functions for which binary operations like addition and multiplication and possibly more are defined (the inverse operations like subtraction and division for example). The binary operations may or may not be *distributive*: $a \times (b + c) = a \times b + a \times c$, *commutative*: $a + b = b + a$ and *associative*: $a + (b + c) = (a + b) + c$. You see that for the algebra of ordinary numbers both the addition and multiplication operations are distributive, commutative and associative (subtraction should be thought of as addition of a negative number $a - b = a + (-b)$, and division by a number as multiplying by the inverse of the number). If you read the next *Math Excursion* you will find that for the algebra of $(n \times n)$ matrices the sum and product are distributive and associative, but whereas matrix addition is commutative, matrix multiplication is not.

A particularly simple algebra we will use in the next chapter is the *Boolean algebra* of binary numbers $\{0, 1\}$. The

algebra is defined by the operations displayed in the table below. They are distributive, commutative and associative.

addition	multiplication
$0+0=0$	$0\times 0=0$
$0+1=1$	$0\times 1=0$
$1+0=1$	$1\times 0=0$
$1+1=0$	$1\times 1=1$

Table A.2: The Boolean algebra.

Algebraic structures that are widely applied in physics are vector spaces, rings, groups and spaces of functions. It turns out that often subjects that begin as pastimes for the mathematically minded end up having great practical use in the realms of science and engineering. What we will see in this book over and again is that in the description of quantum states the notions of vectors, complex numbers, and matrices arise naturally. All these ingredients have a specific underlying algebraic structure. We discuss the algebra of complex numbers in the Math Excursion on page 630, while matrix algebras are described in the next Math Excursion.

Of particular interest in quantum theory is the *algebra of observables* consisting of (hermitian) self-adjoint operators or matrices. These algebras correspond to so-called *Lie algebras*, which are directly linked to the theory of Lie groups, which in turn describe many of the symmetries that play a central role in (quantum) physics. Lie algebras are discussed in more detail on page 634 and Lie groups in the Excursion on page 635.

It is evident that math and physics have co-evolved over centuries leading to a situation where modern theoretical physics makes extensive use of modern and abstract mathematics. It is for that reason that I have decided to throw in more than average math in this semi-popular account of a subject like quantum theory. ♢

♡ On vectors and matrices

The reason for exploring vectors and matrices, is that they play a central role in the mathematical formulation of all of physics and in particular in quantum physics. In classical physics we think of positions, momenta, angular momenta and forces as ordinary three-dimensional vectors. These are *real* vectors because their entries or components are real numbers. In electromagnetism and relativity we have encountered so-called relativistic four-component vectors which are also real. Quantum states are represented by *complex* vectors and physical observables are represented by a class of complex matrices. This excursion highlights some of the more important properties of real vectors and matrices We return to complex vectors and matrices, which play a central role in part II of the book, in a separate *Math Excursion* on page 632.

Real vectors. A vector can be viewed simply as an arrow of a certain length in some n-dimensional Euclidean space $\mathbb{R}^n$. Note that we also have the *null-vector* corresponding to the origin. We denote column vectors by *ket* vectors $|w\rangle$: they are elements of a vector space $\mathcal{V}$, while the row vectors are denoted by so-called *bra* vectors $\langle v|$, and these are elements of a dual vector space $\mathcal{V}^*$. We can add and subtract vectors by just adding or subtracting their corresponding components and scale the vectors by multiplying them by ordinary numbers. These are familiar properties to most of you.

Vector components and choice of basis. If the dimension of the vector space is n, we can choose sets of basis vectors $\{|i\rangle\}$ and $\{\langle i|\}$ and expand vectors as $|v\rangle = \sum_j v^j|j\rangle$ or $\langle v| = \sum_i v_i\langle i|$. You may think of these basis vectors as unit vectors along the different orthogonal axes of the vector space. The reason for this subtle distinction between row and column vectors is that we will encounter different types of vector spaces in this book. We have already seen the example of ordinary Euclidean vectors and

the relativistic Lorentz vectors. The differences between these spaces becomes clear if we look at the definitions for the invariant squared 'length' or the inner product of vectors.

The inner, dot, or scalar product. Having a vector space $\mathcal{V}$ and its dual $\mathcal{V}^*$ we may define an *inner, dot* or *scalar product* between elements $v \in \mathcal{V}^*$ and $w \in \mathcal{V}$ as the number obtained after adding the products of the corresponding entries:

$$\langle v|w\rangle \equiv v \cdot w \equiv \sum_i v_i w^i .$$

As an example we calculate the dot product of two two-dimensional Euclidean vectors:

$$\begin{pmatrix} 2 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = -2 + 1 = -1 .$$

Taking the dot product of a Euclidian vector with itself, $\langle v|v\rangle = |v|^2$ always yields a sum of squares, corresponding to a real number larger or equal zero, which is defined as the *length* of the vector, $|v|$, squared. We also mention that for real vectors the dot product is real and symmetric, $v \cdot w = \langle v|w\rangle = \langle w|v\rangle = w \cdot v$.

As another relevant example we consider the Lorentzian four-momentum vector $p^\mu = (E/c, \mathbf{p})$. The inner product should produce the expression $p_\mu p^\mu = E^2/c^2 - p^2$. This means that the row vector (with lower indices) should be $p_\mu = (E/c, -\mathbf{p})$. It is extremely useful then to define a metric, which is just a matrix $\eta_{ij} = diag(1,-1,-1,-1)$, which maps a column vector to its corresponding row vector like $v_i = \sum_j \eta_{ij} v^j$. And therefore, the inner product can be written using this metric as $v \cdot w \equiv \sum_{ij} g_{ij} v^i w^j$. For the Euclidean case this metric is just the unit matrix $g_{ij} = \delta_{ij} = diag(1,1,\ldots,1)$. Observe that the value of the inner product of a Lorentzian four-vector with itself is not restricted, it can be either positive, negative or zero. Furthermore, if this product is zero, this does not imply that the vector itself has to be zero. It just means that the corresponding particle has vanishing rest-mass.

We have given a graphical representation of the scalar or dot product of two vectors in Figure A.8(a), which underscores the fact that the dot-product produces a number, not a vector, and for that reason it is also called the scalar product.

The exterior or cross product of two vectors. In three dimensions one may indeed also define a 'vector', 'exterior' or 'cross' product between vectors which produces a vector w out of two vectors u and v, and one writes $w = v \times u$. There is no simple extension of such a vector product to general dimensions.

Matrices. Matrices are there in many kinds, appear all over the place and have zillions of applications through the sciences. It refers to a two-dimensional array of elements like for example the apartment building of Figure A.6. The entries of a matrix are often numbers that refer to information about the – in the example at hand – apartment: how many bedrooms, or how many people, or their income, their age etc. In this book we will only employ square $(n \times n)$ matrices that will satisfy various additional properties that derive naturally from certain physical requirements in the specific applications we discuss. There are many ways to look at a matrix: the most neutral way is to say that it is a square array of (real or complex) numbers (see Figure A.7(a)). For example, a distance table between n cities would be like a real $(n \times n)$ matrix. Another way to look at a matrix would be to distinguish the set of diagonal elements, the elements in the upper triangle and the elements of the lower triangle (figure A.7(b)). And sometimes it is convenient to think of a matrix as a stack of n n-dimensional row or column vectors as indicated in Figures A.7(c) and A.7(d).

Matrix algebra. Now the matrices themselves also form a vector space, because we may add and subtract them, there is a 'null-matrix' (with all entries equal zero), and we may multiply a matrix by an arbitrary constant (by just multiplying each entry of the matrix by that constant). There is

Figure A.6: *The Matrix.* A matrix is a two-dimensional array of elements. You may think of this apartment building as a 6×4 matrix, with 6 rows and 4 columns, where the apartments are labeled like the corresponding matrix entries. The entries may refer to information about the inhabitants of the apartments, like the family size, their income, etc. But the analogy is of limited use as we are not adding or multiplying apartment buildings, or assign any meaninbg tot theior aigenvectors and such.. (Source: Alamy.)

more, we may also define a multiplication for matrices as we will see shortly. And in view of the previous *Math Excursion* this means that the set of $n \times n$ matrices form an algebra. To define division for matrices is a little more intricate: we basically define it by multiplying by the inverse of the matrix, where the inverse of A^{-1} of A is defined as the matrix that satisfies $A^{-1}A = AA^{-1} = \mathbf{1}$, where $\mathbf{1}$ is the unit matrix with only ones on the diagonal. This raises the follow-up question of under which conditions the inverse is a well-defined matrix itself. And this question may remind you of the serious elementary school dictum: never divide by the number zero! For matrices the rule is that the inverse exists, if the *determinant* of the matrix is non-zero. This is a number that that can be calculated given the matrix, but we will not go into detail here. Certain matrices have inverses and others have not and there is a relatively simple criterium which tells you if the inverse of a certain square matrix exists. Including the multiplication, we speak of a matrix algebra, as we can perform algebraic manipulations with them similar to what we do with numbers. There is a well-established basic branch of mathematics called 'linear algebra', and there are many textbooks covering the world of matrices in detail.

Matrix as linear transformation of vectors. Now vectors can also be multiplied by matrices to produce another vector, the way that is done is pictorially indicated for a column vector in A.8(b). This action of matrices on vectors is clearly most easily understood if you think of the matrix as a stack of row vectors. The action can also be considered as a *transformation* of a vector into another vector. A simple example may help:

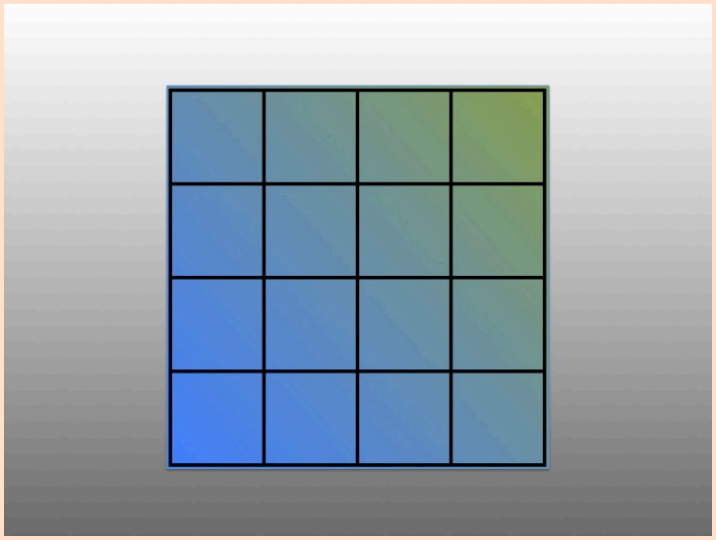
$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 2+1 \\ 2-1 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \end{pmatrix} .$$

The matrix acts as a *linear operator* on the vector space, as it reshuffles the components into linear combinations of them. We may say that $(n \times n)$ matrices map the vector space $\mathcal{V}$ onto itself and we write $A : \mathcal{V} \to \mathcal{V}$. There is for example a particular subset of (3×3) matrices whose action on 'ordinary' vectors corresponds to rotating of those vectors in three-dimensional space R^3.

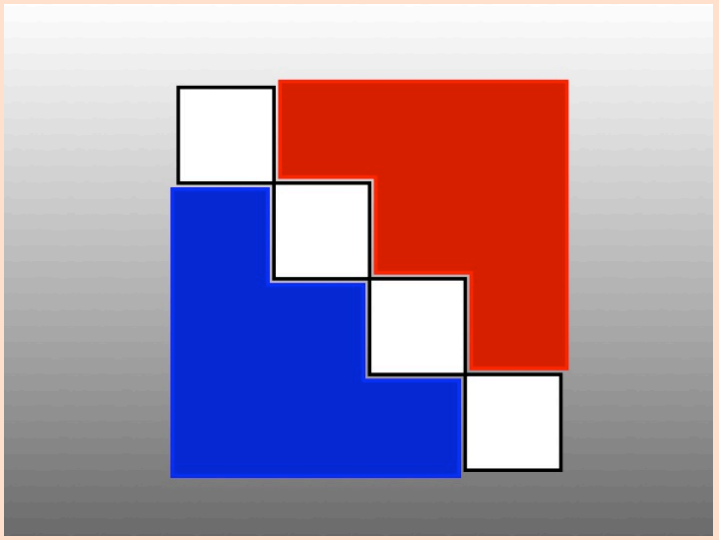
Another example which shows the descriptive power of matrices as operators on state vectors is in (quantum) computation, where generically we think of computation as a sequence of gates, interactions/manipulations or measurements that change the states of a set of (qu)bits.

Such processes or computations can be represented by a product of matrices. Indeed the complete computation is just a matrix mapping the in-state on the out-state vector.

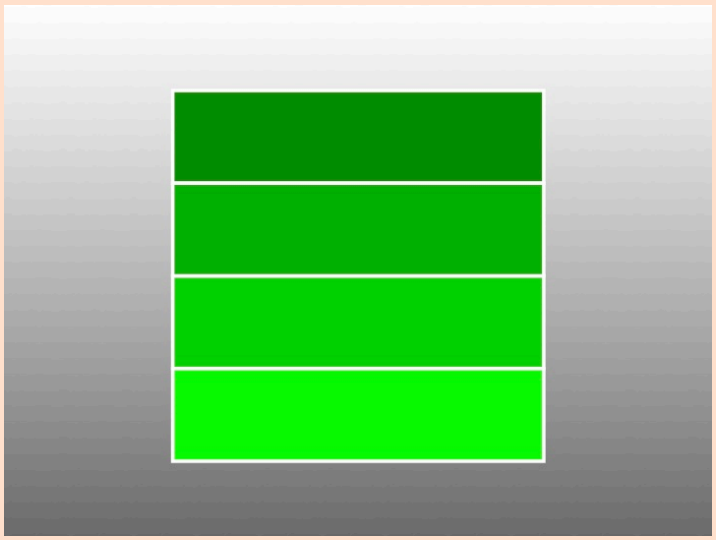
Eigenvectors and eigenvalues. Given a matrix A one defines the *eigenvectors* of A as a set of special vectors



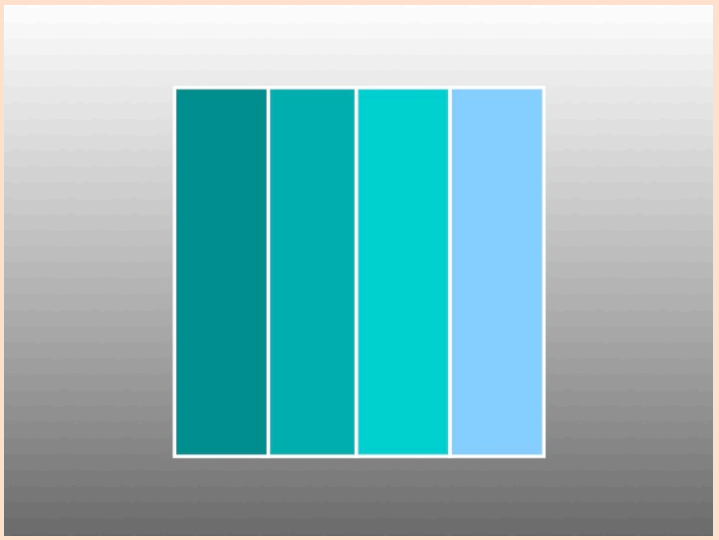
(a) A 4×4 *square matrix* can be thought of as a table of $4^2 = 16$ numbers or symbols representing them.



(b) Square matrix build up of three parts, upper triangular (red), diagonal (white) and lower triangular (blue).

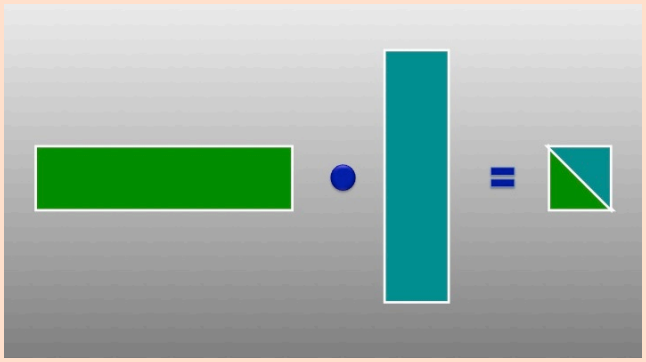


(c) A matrix can also be viewed as a stack of row vectors.

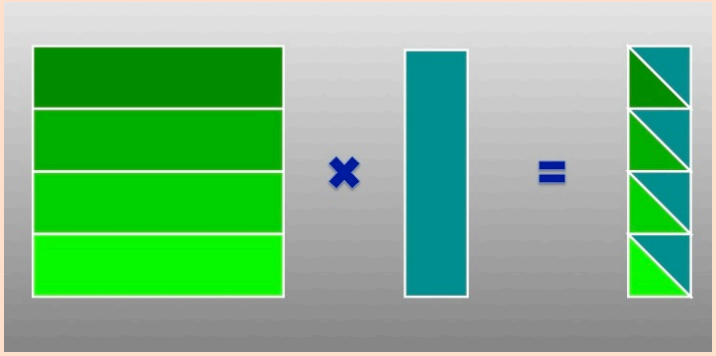


(d) A matrix can also be viewed as a stack of column vectors.

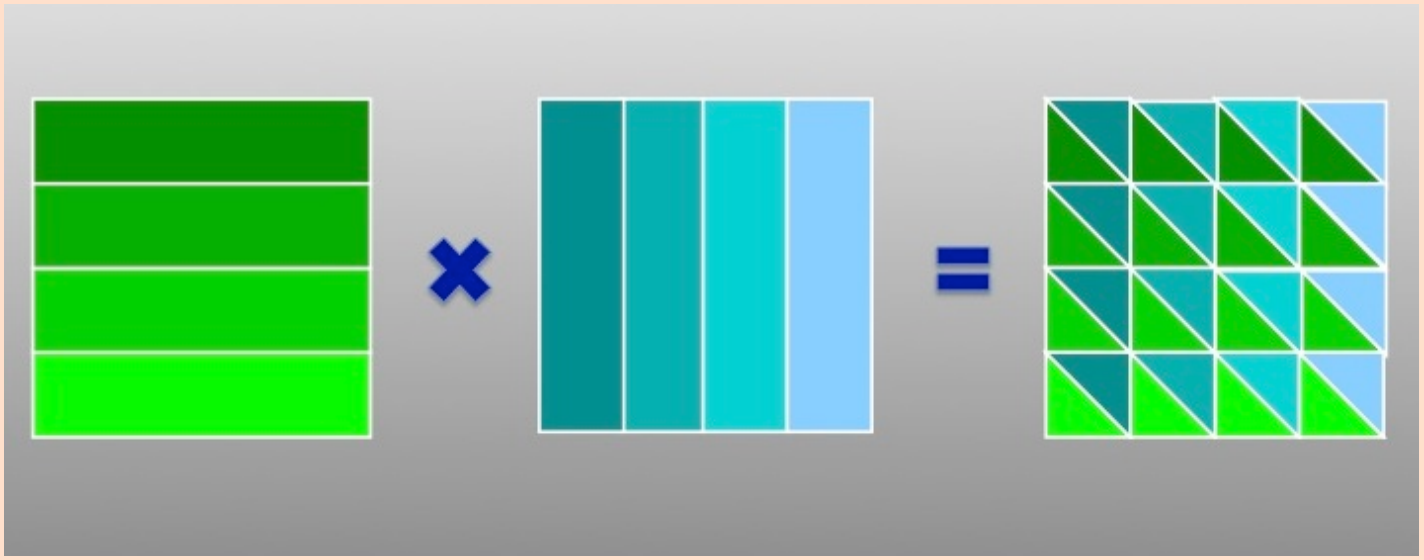
Figure A.7: *Four ways to think about a matrix.* Graphical representation of the many guises of a matrix (artist impression).



(a) The *inner, scalar or dot product* of a row and a column vector yields the single number obtained by adding the product of subsequent row entries with the corresponding column entries: $\langle g|b\rangle = g^* \cdot b = \Sigma_i g_i^* b_i$.



(b) The product of a matrix G with a column vector b yields again a column vector c obtained by taking the dot product of subsequent row vectors of G with the column vector b : $|c\rangle = G|b\rangle = G \cdot b$ meaning $c_i = \Sigma_j G_{ij} b_j$.



(c) The matrix product. Each entry in the product matrix C equals the dot product of the i-th row vector of the first matrix A with the j-th column vector of the second matrix B , so $C_{ij} = \Sigma_k A_{ik} B_{kj}$.

Figure A.8: *Multiplications.* Graphical representation and building up of products of vectors and matrices.

Figure A.9: *Eigenvectors and eigenvalues.* Given a matrix one defines the eigenvectors as a set of special vectors which satisfy an eigenvalue equation (A.8).

$\{|a_k\rangle\}$ that satisfy the following equation:

$$A\,|a_k\rangle = a_k\,|a_k\rangle\ , \tag{A.8}$$

where the numbers a_k are the corresponding *eigenvalues*. So acting on an eigenvector the matrix A gives that same vector back up to a constant, which is by definition the eigenvalue. This is illustrated in Figure A.9. The set of eigenvalues $\{a_k\}$ is called the *spectum* of the matrix. In quantum theory the observables are represented by Hermitean matrices and in that case the eigenvalues are real and the spectrum is called the *sample space* of the operator A.

The matrix product. Once we have defined the action of matrices on vectors the step to the multiplication of matrices is straightforward and we have indicated it in Figure A.8(c). The (ij)-entry of the product matrix $C = AB$ is obtained by the dot product of the $i-th$ row vectors of A with the $j-th$ column vector of B. Let us again give a simple example:

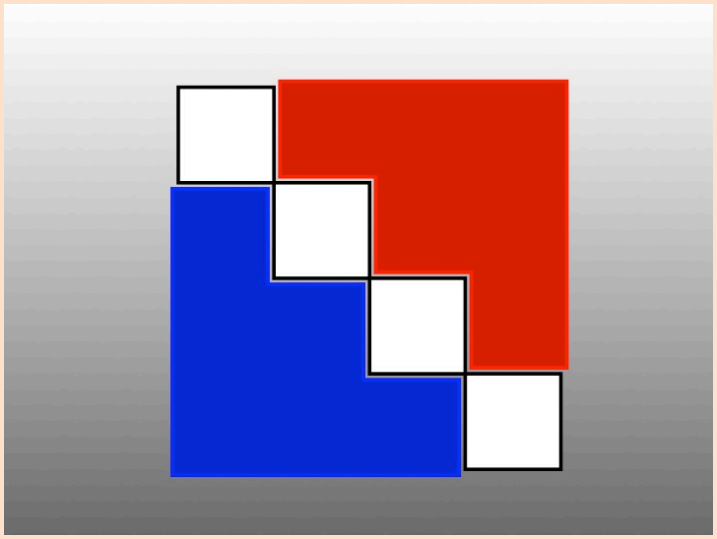
$$\begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 3 & -1 \\ -1 & -3 \end{pmatrix}\ . \tag{A.9}$$

Types of matrices. A distance table between n different cities is a square $(n \times n)$ matrix, a rather special one for sure, because its diagonal elements are all zero and it is symmetric with respect to that diagonal: the upper diagonal and lower diagonal matrices are each other's mirror image. Such a matrix is completely determined by specifying its $n(n-1)/2$ upper triangular entries.

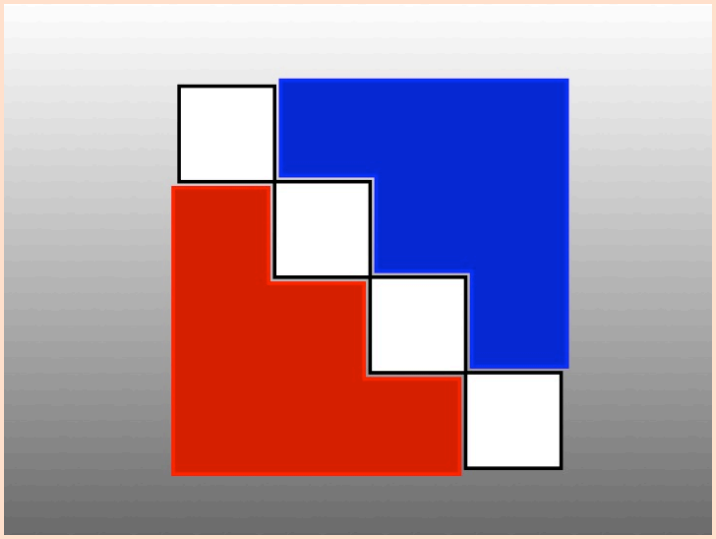
Depending on the situation we may want to put additional constraints which define a subset of matrices. If the additional properties are preserved under the basic matrix operations, the subset forms a subalgebra of the original algebra. The additional properties involve typical matrix manipulations which we have represented symbolically in figure A.10. A fundamental notion is the *transpose* of a matrix denoted by the matrix A^{tr}, which is obtained from A as indicated in Figures A.10(a) and A.10(b), written in terms of its entries one has $(A^{tr})_{ij} \equiv A_{ji}$. The transpose can be obtained by mirroring the matrix in the diagonal but can also be obtained by interchanging rows and columns. Repeating the operation brings you back to the original matrix. What happens if we take the transpose of a product of matrices? Referring again to Figure A.8(a), one sees that taking the transpose of matrix $C = AB$ on the right-hand side we get a matrix which is the product of the transposes, but in the opposite order: $C^{tr} = B^{tr}A^{tr}$.

Now it is also straightforward to define a symmetric or antisymmetric matrix as the ones that satisfies $A = \pm A^{tr}$ (see Figure A.10(c)). Note that a symmetric $(n \times n)$ matrix contains $n(n+1)/2$ real numbers, while the antisymmetric one has only $n(n-1)/2$, because the diagonal elements must be zero for the latter.

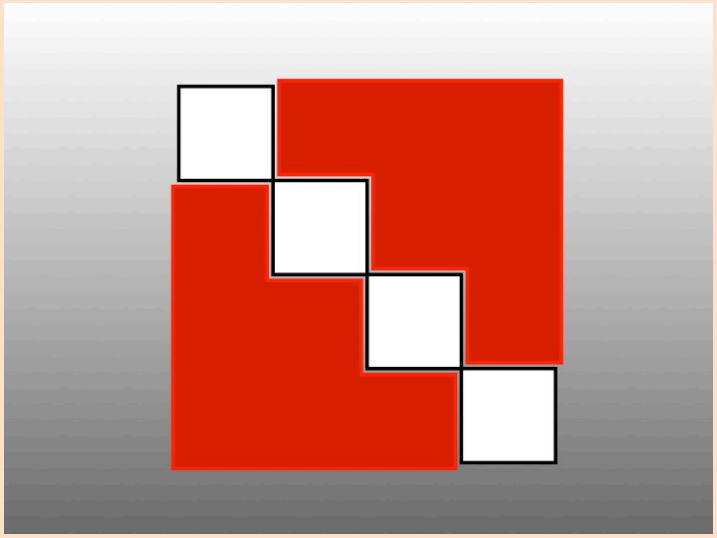
Invariance of the inner product. We have shown that the product of a vector with itself defines the length of a vector, and we all know that the length of a vector does not change if we rotate the vector around. So we say that the length of a vector is *invariant* under rotations. Also the angle between two vectors is invariant under rotations. In other words, the inner product of two vectors is invariant under rotations.



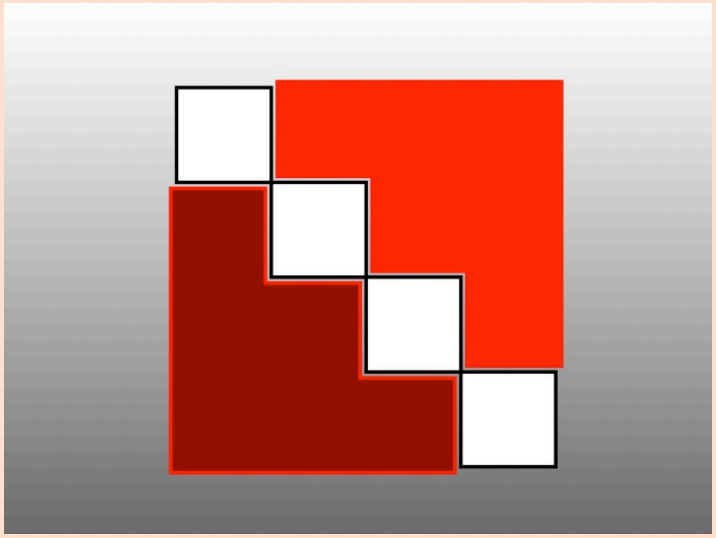
(a) A *square* matrix build up of three parts, upper triangular (red), diagonal (white) and lower triangular (blue).



(b) *Transpose* of the matrix depicted in (a), obtained by reflecting in the diagonal, or by interchanging the rows and columns of the matrix.



(c) A *symmetric* matrix is equal to its transpose. A distance table between four cities would be a symmetric matrix with zeros along the diagonal.



(d) An *antisymmetric* matrix is a matrix whose transpose equals minus that matrix. In other words: $A_{ji} = -A_{ij}$.

Figure A.10: *Matrix properties.* Graphical representation of some basic properties of matrices.

As the rotations involve a transformation of the vector into another vector, it follows that rotations can be represented by matrices acting on the vector space $\mathcal{V}$. And for real vectors this matrix has to be a real matrix. Imagine we act with a rotation matrix R on $|v\rangle$. We may write $|v'\rangle = R|v\rangle$, and it then follows that $\langle v'| = \langle v| R^{tr}$. Invariance of the inner product of two arbitrary vectors now requires that

$$\langle v'|w'\rangle = \langle v| R^{tr} R |w\rangle = \langle v|w\rangle \quad \Rightarrow \quad R^{tr} = R^{-1}. \quad \text{(A.10)}$$

What this equation is telling us is that the matrices R that represent rotations must satisfy the property that their transpose equals their inverse. Matrices that have that property are called *orthogonal* matrices. There is an additional important property that these matrices must satisfy. If you realize that if we do two subsequent rotations on a vector, then that is the same as doing a single rotation that brings the vector directly from its original to its final orientation. Translated in the language of rotation matrices this means that the product of two orthogonal matrices is again an orthogonal matrix. And one says that the collection of all such matrices define a *group*, for the case at hand this is the so-called rotation group in n-dimensions denoted by $SO(n)$. The $SO(n)$ group has $n(n-1)/2$ independent elements.

What about the four-vectors whose inner product involves not the unit matrix, but rather the diagonal 4×4 matrix $\eta_{\mu\nu} = diag(1,-1,-1,-1)$? Now we must impose a different invariance condition on the transformation matrices Λ, it reads $\Lambda^{tr}\ \eta\ \Lambda = \eta$. The Lorentz transformations are defined by the condition that they leave the inner product matrix or metric, η, invariant. The associated, so-called Lorentz group is then denoted as $SO(1,3)$, as the metric has one plus sign and three minus signs. ♡

♠ On vector calculus

In this excursion we touch on three important theorems with respect to integrating equations involving the vector derivative ∇ of fields. These theorems refer respectively to the line integral, an integral over an area and a volume integral.

Operators involving the vector derivative ∇. 
We have been talking about fields such as a force field $\mathbf{F}(x)$, a current density $\rho(x)$ or the electric and magnetic fields $\mathbf{E}(x)$ and $\mathbf{B}(x)$. Such a vector field defines a vector at any point in space(time). We have also encountered the vector of derivatives called *nabla*:

$$\nabla = (\partial_{x_1}, \partial_{x_2}, \partial_{x_3}),$$

which plays a fundamental role in the calculus of (vector) fields which features as we have seen in the Maxwell equations of electromagnetism, but as a matter of fact it plays an equally important role in the subject of fluid dynamics. If the equations involve the nabla operator, then solving the equation means that we somehow have to 'integrate' the equation. The mathematics involved is denoted as *vector calculus* in contradistinction to *vector algebra*, which only involves algebraic manipulations of vectors.

The **gradient** of a scalar function yields a vector field. In this chapter we have encountered various definitions where a vector field was defined as the vector derivative or *gradient* of a scalar potential function $V(x)$, like for example the relations:

$$\mathbf{F}(\mathbf{x}) = -\nabla V(x),$$
$$\mathbf{E}(\mathbf{x}) = -\nabla V(x).$$

When discussing the Maxwell equations we also encountered vector derivatives of vector functions. Here we distinguish the following two possibilities:

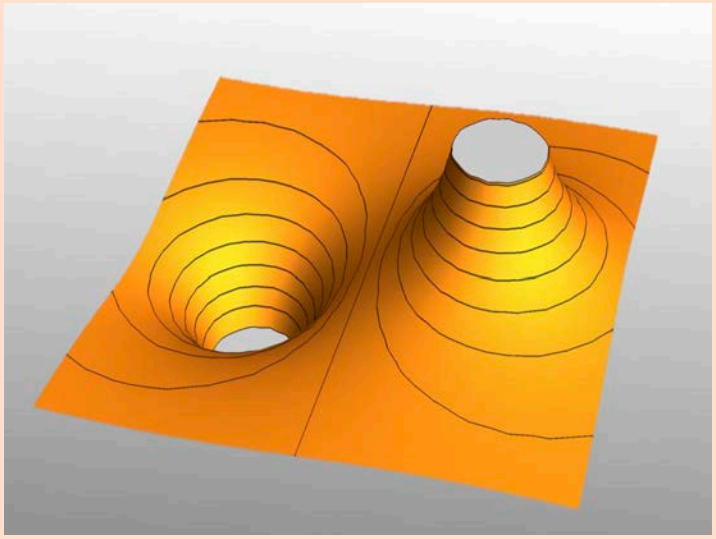


Figure A.11: *The electrostatic potential for a dipole.* This is the potential $V(x,y)$, with some equipotential lines, resulting from two opposite charges placed placed at opposite points on the x axis.

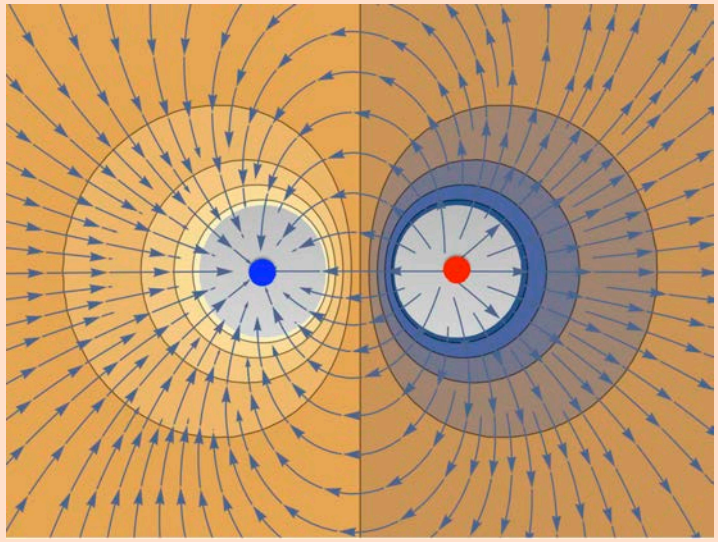


Figure A.12: *The electric dipole field.* This is the dipole field $\mathbf{E}(x,y)$ corresponding to minus the gradient of the potential depicted in the previous figure. We have drawn the *field lines*; these are the stream lines of the field. At any point the field is directed along the tangent of the line going through that point, and the magnitude is proportional to the density of lines around that point. The closed *equipotential lines* are projected in the plane, and we see that the field lines are orthogonal to them. This means that the field lines are the projections of the lines of *steepest descent* on the surface of the previous figure.

(i) The **divergence** of a vector field, which yields a scalar function, for example:

$$\rho(x) = \boldsymbol{\nabla} \cdot \mathbf{E}(\mathbf{x})\,.$$

(ii) The **curl** of a vector field, which yields another vector field, for example:

$$\begin{aligned} \mathbf{j} &= \boldsymbol{\nabla} \times \mathbf{B}\,, \\ \mathbf{B} &= \boldsymbol{\nabla} \times \mathbf{A}\,. \end{aligned}$$

These operations contain first-order derivatives and are thus linear in nabla. We also need higher-order derivatives, apart from definitions like the 'Laplacian' $\Delta \equiv (\boldsymbol{\nabla} \cdot \boldsymbol{\nabla})$, there exist additional mathematical identities. In Chapter I.2 we used already two of them:

$$\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \mathbf{A}) = 0\,, \tag{A.11a}$$

$$\boldsymbol{\nabla} \times (\boldsymbol{\nabla} V) = 0\,. \tag{A.11b}$$

One more useful identity is basically rewriting the repeated vector product of the nabla operator:

$$\boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \mathbf{A}) = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \mathbf{A}) - (\boldsymbol{\nabla} \cdot \boldsymbol{\nabla})\mathbf{A}\,, \tag{A.12}$$

where the Laplacian in the last term is understood as acting on the components of vector $\mathbf{A}$ individually.

We emphasize that the above are identities, meaning that they hold for any vector field $\mathbf{A}(\mathbf{x},t)$ and any scalar field $V(\mathbf{x},t)$.

To solve systems like the Maxwell equations we are interested in 'integrating' expressions involving the basic vector derivatives, this is facilitated by some powerful theorems that we will look at next.

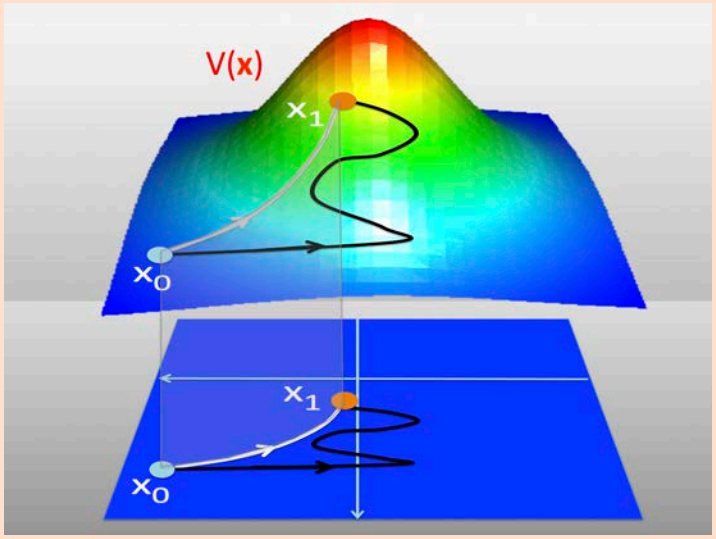


Figure A.13: *A line integral.* In the upper picture we give a two-dimensional potential surface $V(\mathbf{x})$. The force field is defined as $\mathbf{F}(\mathbf{x}) = -\boldsymbol{\nabla} V(\mathbf{x})$. If we choose a path from point $\mathbf{x}_0$ to $\mathbf{x}_1$, we can integrate $\mathbf{F}$ along that path, meaning that we integrate the component tangential to the path. This line integral yields the value $W = V(\mathbf{x}_0) - V(\mathbf{x}_1)$ which equals the work performed by the force, which in this is negative. We had to perform a force to go uphill and therefore, the potential energy was increased. Note that the outcome is *independent* of the path chosen.

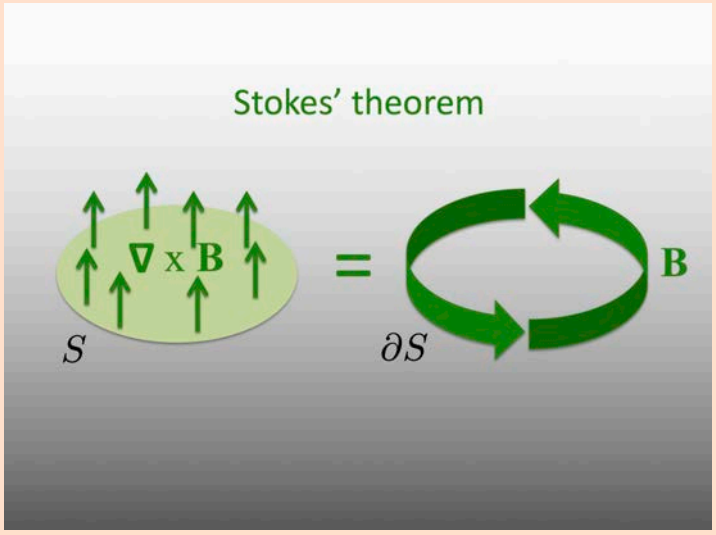


Figure A.14: *A surface integral.* The figure is a pictorial representation of Stokes' law, which says that integrating the component of the curl of a vector field ($\boldsymbol{\nabla} \times \mathbf{B}$) orthogonal to an arbitrary surface, over an area A, equals the line integral of that vector field along the closed boundary contour ∂A of that area.

Integration theorems for vector derivatives. 

We have seen that the Maxwell equations are first-order partial differential equations for the vector fields $\mathbf{E}$ and $\mathbf{B}$. That means that given the sources one could solve these equations by integrating them. It is here that some powerful integration theorems for vector derivatives can be exploited. These lead to what is often called the integrated form of the Maxwell equations, which no longer contain any spatial derivatives of the fields.

We will consider the following cases:

(i) The *line integral* of a gradient field along a curve γ, for example:

$$\int_{\mathbf{x}_0}^{\mathbf{x}_1} \mathbf{F}(\mathbf{x}) \cdot \mathbf{dl} = -\int_{\mathbf{x}_0}^{\mathbf{x}_1} \boldsymbol{\nabla} V(\mathbf{x}) \mathrm{d}\mathbf{x} = V(\mathbf{x}_0) - V(\mathbf{x}_1)\,,$$

where the line element $\mathbf{dl}$ is the unit vector tangent to the curve. We discussed this example already in Chapter I.1. In ordinary language this refers to the statement that if you apply a force on an object, then the integral of that force along a given path corresponds to the work applied to the object and that equals the increase of the potential energy of the object, as we have indicated in Figure A.13. This increase equals the difference of the potential energies at the endpoints of the path. The fact that the difference only depends on the endpoints means that the increase of energy is *not* dependent on the path chosen. If you want to climb to the top of a mountain you can choose between a path that is long and not so steep or a very short very steep path in either case you must deliver the same amount of energy.

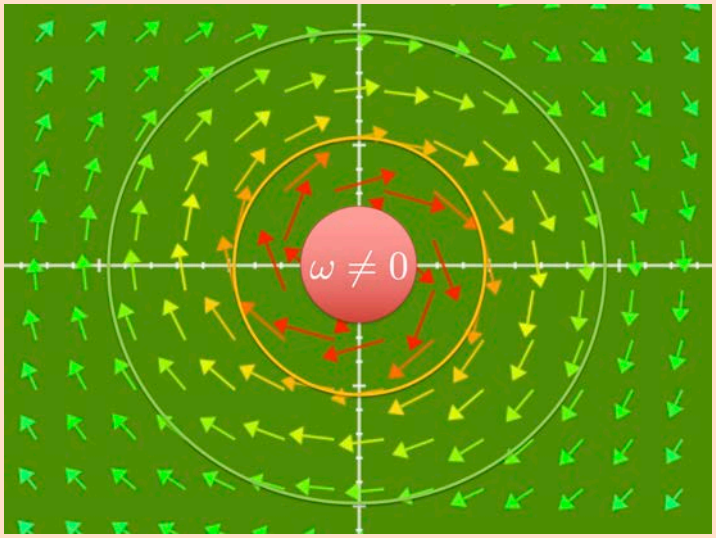


Figure A.15: *A vortex field.* The velocity field $\mathbf{v}(\mathbf{x})$ of an ideal or free vortex around a source where the vorticity $\boldsymbol{\omega}$ is non-zero in a small region around the origin and pointing along the axis perpendicular into the plane of the figure.

Figure A.16: *A tornado.* A tornado is an aerodynamical flow pattern with vorticity and a non-zero circulation.

(ii) The *surface integral of a curl* over a given area A, known as Stokes' theorem:

$$\int_A \boldsymbol{\nabla} \times \mathbf{B} \cdot \hat{\mathbf{n}}\, d^2S = \oint_{\partial A} \mathbf{B} \cdot \mathbf{dx}\,,$$

where on the left-hand side $\hat{\mathbf{n}}$ is the unit vector perpendicular to the surface element d^2S, and on the right-hand side we integrate the vector field $\mathbf{B}$ along the boundary ∂A of the surface area. This mathematical theorem is illustrated in Figure A.14.

The most familiar application is in fluid mechanics where the vector field defining the flow is the velocity field $\mathbf{v}(\mathbf{x}, t)$. The *vorticity* $\boldsymbol{\omega}$ of the fluid is then defined as the curl of the velocity field:

$$\boldsymbol{\omega} = \boldsymbol{\nabla} \times \mathbf{v}\,.$$

The simplest example is a situation where the vorticity to be non-zero only on the $z-$axis, as a constant vector in the positive $z-$direction. Then the solution for the velocity field is the familiar cylindrical free vortex flow around the z axis, corresponding to an ideal vortex. A related quantity is now the *circulation* of the flow as a surface integral of the vorticity, which then equals the line integral of the velocity around a closed loop bounding surface area. In the example where $\boldsymbol{\omega} = k\hat{\mathbf{z}}$ only on the $z-$axis, one obtains that for a loop winding once around the $z-$axis, the circulation γ equals $\gamma = n\,k$. Taking a horizontal circle around the z-axis we get a cylindrically symmetric, *free vortex field* with an angular velocity that drops off inversely proportional with the radius: $v(r) = k/2\pi r$, as depicted in Figure A.15. A beautiful, not so ideal vortex is the tornado depicted in Figure A.16.

In electrodynamics one applies Stoke's theorem to Ampère's law yielding

$$\oint_{\partial A} \mathbf{B} \cdot \mathbf{dx} = \int \mathbf{j} \cdot \hat{\mathbf{n}}\, d^2S\,.$$

This is basically the 'integrated form' of Ampère's law, the equation $\boldsymbol{\nabla} \times \mathbf{B} = \mathbf{j}$, that was already depicted on the left in Figure I.1.18.

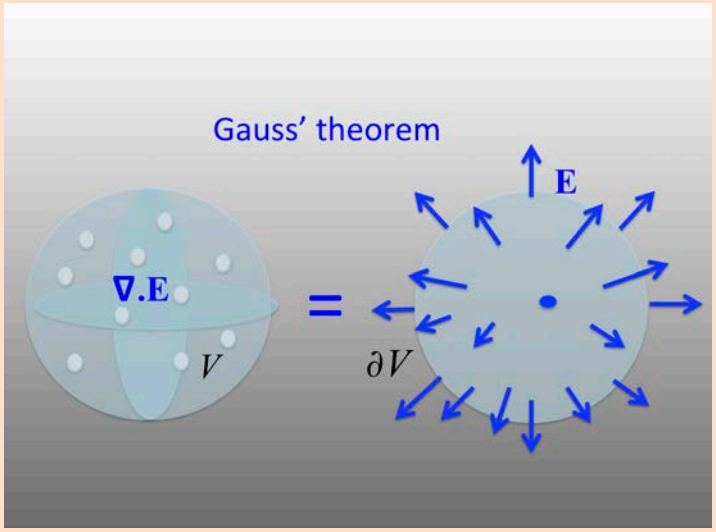


Figure A.17: *A volume integral.* The figure illustrates Gauss' law states that the volume integral of the divergence ($\boldsymbol{\nabla} \cdot \mathbf{E}$) of a vector field $\mathbf{E}$ equals the surface integral of the perpendicular component of that vector field over the closed surface (∂V) bounding the volume V.

Stoke's theorem also applies to the magnetic flux through a bounded surface, which becomes equal to the loop integral of the vector potential $\mathbf{A}$, which is defined by the equation $\mathbf{B} = \boldsymbol{\nabla} \times \mathbf{A}$:

$$\Phi = \int \mathbf{B} \cdot \hat{\mathbf{n}}\, \mathrm{d}^2S = \oint_{\partial S} \mathbf{A} \cdot \mathbf{dx}\,.$$

(iii) The *volume integral of a divergence* over volume V known as Gauss' theorem:

$$\int \boldsymbol{\nabla} \cdot \mathbf{E}(\mathbf{x}) \mathrm{d}^3V = \int_{\partial V} \mathbf{E} \cdot \hat{\mathbf{n}}\, \mathrm{d}^2S\,,$$

where the integral on the right-hand side is over the closed surface S bounding the volume V. This theorem is depicted in Figure A.17.

We can apply it to the first Maxwell equation I.1.26 as follows:

$$\int_V \rho(x) \mathrm{d}^3V \;=\; \int_{\partial V} \mathbf{E} \cdot \hat{\mathbf{n}}\, \mathrm{d}^2S \;=\; \mathrm{Q}\,,$$

telling us that integrating the perpendicular component of the electric field over a closed surface bounding a volume yields the total electric charge inside that volume. ♠

♣ On probability and statistics

> . . . But ignorance of the different causes involved in the production of events, as well as their complexity, taken together with the imperfection of analysis, prevent our reaching the same certainty [as in astronomy] about the vast majority of phenomena. Thus there are things that are uncertain for us, things more or less probable, and we seek to compensate for the impossibility of knowing them by determining their different degrees of likelihood. So it is that we owe to the weakness of the human mind one of the most delicate and ingenious of mathematical theories, the science of chance or probability.
>
> (Laplace, 1889)

Probabilities. A variable x can take on values, in a discrete or maybe a continuous set, a *domain* or a *sample space* we will denote by $\mathcal{X} = \{x_i\}$. A *random* or *stochastic variable* is one where we associate with that variable a *probability distribution* over the domain, so we introduce a probability function $p_i = p(x_i)$ that gives the chance or probability that x will have the value x_i. As the variable x always carries some value, we must require that the probabilities add up to one:

$$\Sigma_i \, p_i = 1 \, . \tag{A.13}$$

Given a random variable and its probability distribution, we can calculate the average outcome of a number of statistically independent measurements of x or for that matter any function $f(x)$ of x. It is simply given by the *expectation value* or *average* defined as:

$$< f >= \Sigma_i \, p_i f(x_i) \, . \tag{A.14}$$

So for a fair dice we have that $\mathcal{X} = \{1, 2, \ldots, 6\}$ and $p_i = 1/6$ for all i, and therefore one calculates for example that $< x >= \frac{1}{6}\Sigma_i \, i = 7/2$ and $< x^2 >= \frac{1}{6}\Sigma_i \, i^2 = 91/6$.
We can ask the same questions for the sum outcomes if we throw two dice, we have now to first determine the domain of $x = x(1) + x(2)$ to obtain $\{2, 3, \ldots, 12\}$. The probability for each outcome equals the number of distinct combinations for the two dice to get the given answer. For example, from the $6 \times 6 = 36$ possible combinations, the outcome $x = 7$ can be obtained in 6 distinct ways, namely,

$$(x(1), x(2)) = (1,6), (2,5), (3,4), (4,3), (5,2), (6,1) \, .$$

So, the probability $p(x = 7) = 6/36 = 1/6$. One can similarly construct distributions $P(x, n)$ for n dice, and these are depicted in Figure A.18 for an increasing number of dice.

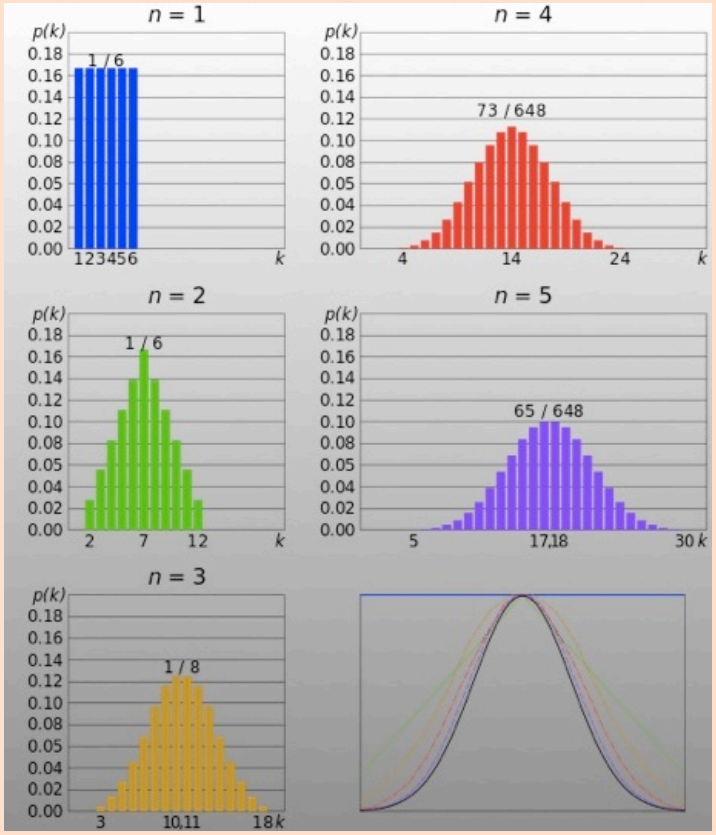


Figure A.18: *The distributions* $P(x, n, 6)$, with $x = x(1) + \ldots + x(n)$ for throwing n fair dice. For large n this symmetric distribution approaches the normal or Gaussian distribution.

Another important quantitative measure of a distribution is the *standard deviation* σ and its square, called the *variance* or *mean square deviation*, which is defined as:

$$\sigma^2 = < (x - < x >)^2 > = < x^2 > - < x >^2 \ . \qquad \text{(A.15)}$$

The variance is a measure of the width of the distribution. For the dice examples one finds that for one dice $\sigma = \sqrt{35/12} = 1.71$ and for the pair $\sigma = \sqrt{35/6} = 2.42$.

Statistics. Having a stochastic variable one can make measurements at a series of times t_m, and one may study the frequency distribution of outcomes and compare it for example with a theoretically predicted probability distribution. Here we enter the field of statistics, of statistical analysis. The challenge of statistical analysis is to understand from the measurements, what the set of sample values you have taken tells you about the true distribution. The central and vital question is what conclusions you can draw from some experiment and with what degree of certainty or confidence.

Say the length of males in *cm* for a certain country has a certain distribution $H(h)$, which may peak around $170\ cm$. Now we can take a sample of the population and from the sample construct the sample distribution, which now is like an approximation of the real distribution, and it will not surprise you that by making the sample ever larger the approximation will get better. It may also be that you are probing a space of choices that people make and try to predict the probability of the next choices that will be made. The business of polling is in this category. Politicians and public media frequently demonstrate their ignorance where it comes to understanding statistics, and sometimes proudly so. In science, however, we must insist on a solid understanding of statistics to interpret what we see, or think to see, and in order to draw balanced and reliable conclusions, taking the uncertainties which are always there, properly into account.

Central limit theorem. Often one is interested in a quantity y, which is dependent on many different independent random variables. The height of people for example may be written as the sum of other random variables $x^{(m)}$ with $m = 1, \ldots, M$, where each may have its own distribution $p(x^{(m)})$. Under general conditions on the distributions $p(x^{(m)})$ the distribution $P(y)$ we are interested in will approach the *Gaussian* or *normal distribution*. So, quantities that equal the sum of many random variables, which need *not* be normally distributed themselves, tend to be normally distributed! This is as true for the velocity distribution of particles in a gas kept at a given temperature, as it is for the height distributions in a population, or for the frequency of errors, but also for the minimal uncertainty wave packet describing a quantum particle. The importance of this normal distribution cannot be overstated as it pops up in any serious field of study. This is nicely expressed in the following quote of Sir Francis Galton, the Victorian progressive, polymath, statistician, sociologist, psychologist, anthropologist, eugenicist, tropical explorer, geographer, inventor, meteorologist, proto-geneticist, and psychometrician:

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the 'law of frequency of error' [the normal or Gaussian distribution]. Whenever a large sample of chaotic elements is taken in hand and marshalled in the order of their magnitude, this unexpected and most beautiful form of regularity proves to have been latent all along. The law . . . reigns with serenity and complete self-effacement amidst the wildest confusion. The larger the mob and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of unreason.
>
> (Galton, 1889)

The normal distribution depends on two parameters, its *mean* or *expectation* μ and its *variance* σ^2, and it is given

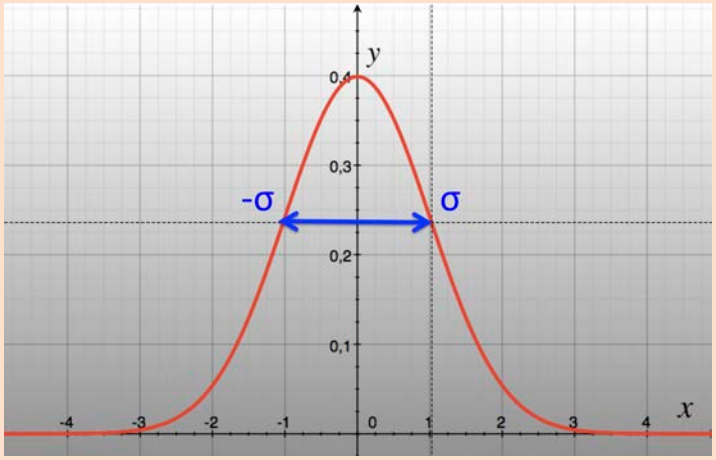


Figure A.19: *The Gaussian or normal distribution,* with variance $\sigma^2 = 1$ and mean $\mu = 0$.

by the following expression,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} . \qquad (A.16)$$

We have depicted the normal distribution in Figure A.19 with its familiar bell shape.

Statistical physics. To describe the macroscopic properties of systems like gases, fluids, plasmas one does not need to know the precise properties of all individual particles making up the system. Fortunately, because that would amount to solving some 10^{23} coupled partial differential equations. If we put the particles say in a container, then each of the particles has a well-defined phase space that is the same for all of them, but each particle may sit in a different corner of the phase space. Boltzmann made the assumption that such a macro-system may then be characterized by some distribution of the particles over phase space.

For a simple gas or fluid, he introduced the distribution function $f(\mathbf{x}, \mathbf{v}, t)$, giving the probability density for a particle in the gas to have position $\mathbf{x}$ and velocity $\mathbf{v}$ at time t. This function will have some generic features. He in fact showed that this distribution function had to satisfy some fundamental equation which now carries his name. From f one can derive the number density distribution, $n(\mathbf{x}, t) = \int f(\mathbf{x}, \mathbf{v}, t)\, d^3\mathbf{v}$.

If the system is in equilibrium, one has that the distribution f is time independent. In a gas in equilibrium (without external forces) we expect the particles to spread out evenly over the volume, so f will also be $\mathbf{x}$ independent, and because of the interactions one expects that the energy will be quite equally distributed over the particles. If we keep the gas at a fixed temperature, so that the average energy per particle equals $3kT/2$, this leads to the well-known Maxwell-Boltzmann equilibrium velocity distribution:

$$f(\mathbf{v}; T) = \left(\frac{m}{2\pi kT}\right)^{3/2} e^{-m|\mathbf{v}|^2/2kT} , \qquad (A.17)$$

which is a 3-dimensional Gaussian distribution.

Entropy. With a given distribution p, one can always associate a certain Gibbs-Shannon or *information entropy* $S(p)$ with,

$$S(p) = -\Sigma_i p_i \log_2 p_i . \qquad (A.18)$$

The entropy is thus a number that you can calculate given a distribution. If the outcome is certain, then one has for one particular i that $p_i = 1$ while the others are zero, and one finds that $S = 0$. On the other hand if the outcome is maximally uncertain we will have that N states $p_i = 1/N$ for all i, implying that the entropy will attain its maximal value $S = \log_2 N$. Another interesting property is that entropy is an additive quantity, if one combines two independent distributions. Imagine throwing simultaneously a fair coin and a fair dice with distributions $p^{(1)}$ and $p^{(2)}$, then there are $2 \times 6 = 12$ states with a combined distribution $p = p^{(1)} \times p^{(2)}$. The entropies then satisfy the additive relation: $S = S^{(1)} + S^{(2)}$. In other words, if one finds in an experiment that the additive property does not hold this indicates some interdependence between the variables, which in physical terms means that the two components of the system interact. It is therefore certainly pos-

sible to have a closed system consisting of two interacting subsystems, where the entropy of one subsystem actually decreases, as long as the entropy of the other subsystem increases by an equal or larger amount, as to make sure that the whole system satisfies the second law. For example, if one has a mixture of different particle types, which at some point will start binding, the bound state represents a lower energy state, and thus in this transition heat will be released, which corresponds to pure entropy production. Here we see that on the one hand the interactions cause more structure, a higher level of order and thus less entropy in the particle component of the system, but at the same time the entropy of the system as a whole will increase because of the amount of heat that is produced.

Maximal entropy principle. If you have a certain sample space, you may want to consider different distributions $p^{(m)}$ over that space and compare their entropies. Then an interesting fact is that the distribution that maximizes the entropy over the set of distributions $\{p^{(m)}\}$ is the best guess you can make, assuming that you know nothing else about the process or the distribution you are studying except that the probabilities add up to one. But in many cases you do know more, for example you know the average outcome of some observable $\lambda(x)$, so $<\lambda(x)>=\lambda_0$. Then you want to maximize the entropy under the additional constraint that $<\lambda(x)>=\sum_i p_i\lambda(x_i)=\lambda_0$, and that will lead to another maximal entropy distribution. So the maximal entropy distribution is the least biased probability distribution under the given set of constraints. Many of the distributions that play an important role in nature are maximal entropy distributions. Let us look at some of the familiar cases:

(i) We define the information entropy $H(\{p_i\};\{\lambda_k\})$ as the entropy but with the constraints added with a parameter λ_k. The trivial case is where we impose that the sum of the probabilities equals one:

$$H(p_i;\lambda_k) = -\sum_i p_i \ln p_i - \lambda_0(\sum_i p_i - 1)\,. \qquad \text{(A.19)}$$

We maximize H with respect to the $\{p_i\}$ and $\{\lambda_k\}$ by requiring the partial derivatives to be zero:

$$\left(\frac{\partial H}{\partial p_i}\right) = \quad -\ln p_i - 11 - \lambda_0 = 0\,, \qquad \text{(A.20)}$$

$$-\left(\frac{\partial H}{\partial \lambda_0}\right) = \quad \sum p_i - 1 = 0\,. \qquad \text{(A.21)}$$

The first equation yields that p_i is constant $p_i = p$; substitution in the second equation yields $Np - 1 = 0$, so that $p = 1/N$, corresponding to the well-known case of fixed energy or the micro-canonical ensemble.

(ii) Let us now take a continuous energy type distribution where we know the average energy to be ε^*. Then we must add to the expression (A.19) the constraint term $-\lambda_1(\int_0^\infty p_i\varepsilon - \varepsilon^*)$, yielding for the first equation:

$$-\ln p - 1 - \lambda_0 - \lambda\varepsilon\,, \qquad \text{(A.22)}$$

with solution

$$p(\varepsilon) = Ce^{-\lambda\varepsilon}\,.$$

From the first constraint we get:

$$\int p(\varepsilon)\,d\varepsilon = C(-\frac{1}{\lambda})\,e^{-\lambda\varepsilon}\Big]_0^\infty = \frac{C}{\lambda} = 1\,, \qquad \text{(A.23)}$$

so we learn that $C = \lambda$. Substitution in the second constraint yields another relation that we can solve for both parameters:

$$C\int \varepsilon\, e^{-\lambda\varepsilon}\,d\varepsilon = \varepsilon^*\,. \qquad \text{(A.24)}$$

Let us rewrite

$$-C\frac{d}{d\lambda}\int e^{-\lambda\varepsilon}\,d\varepsilon = -C\frac{d}{d\lambda}(\frac{1}{\lambda}) = \frac{C}{\lambda^2} = \varepsilon^*\,, \qquad \text{(A.25)}$$

which yields $C = \lambda = 1/\varepsilon^*$ and we obtain the simple exponential distribution:

$$p(\varepsilon) = \frac{1}{\varepsilon^*} e^{-\varepsilon/\varepsilon^*} . \qquad (A.26)$$

(iii) A similar calculation can be set up for the case where we have prior knowledge about the variance of the distribution, in which case one obtains a Gaussian distribution, like the celebrated Maxwell-Boltzmann distribution.

The maximal entropy principle is a powerful tool for constructing the optimal distribution satisfying a certain number of constraints. And we see that this is completely consistent with our discussion of statistical mechanics in chapter I.1. A virtue of the maximal entropy principle is that it nicely separates the purely statistical and the more physical aspects in the approach to macroscopic systems. This approach to statistical mechanics, inspired by the work of Gibbs and Shannon, was introduced in 1957 by the American physicist Edwin Thompson Jaynes.

Quantum entropy. In quantum theory, probability plays an important role even if we consider a system consisting of a single particle, as its wave function or state vector is a probability amplitude that encodes the probability for obtaining certain outcomes of measurements of an observable. Therefore, probability is built in right from the start for any quantum system. and you expect that there is some meaning to the notion of entropy as well. Indeed, there is, the quantum entropy was defined by Von Neumann much in parallel with its classical precursor:

$$S = -\mathrm{Tr}\ \rho \log \rho . \qquad (A.27)$$

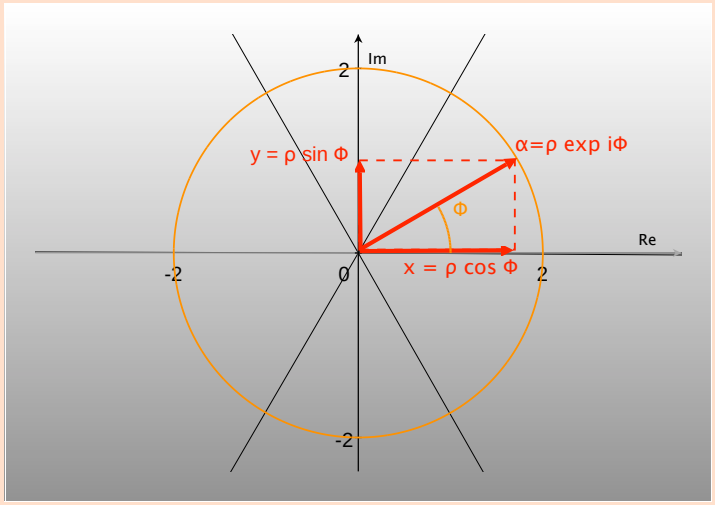
In this expression, ρ is the so-called density matrix of the system as discussed in Chapter II.1, which represents the state of the system. The symbol Tr stands for the trace of a matrix, which equals the sum of its diagonal components. The Von Neumann entropy is a measure for the degree of *entanglement* of a multicomponent quantum system. ♣

♠ On complex numbers

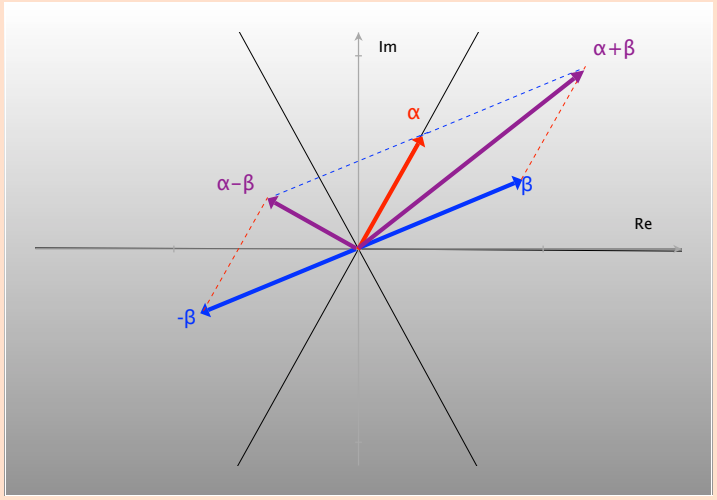
> Mathematics is one of the few places where complexification often stands for simplification.

Number systems. It is interesting to note how number systems have been extended through history. A natural starting point are the *natural numbers* or positive integers, and we know how to add and subtract them, where to stay within the set of natural numbers the subtraction is restricted to numbers smaller (or equal if we include zero in the set). We can extend the definition of subtraction to all natural numbers but that forces us to augment the set with the exquisite number 'zero' and the negative integers. One defines multiplication as an operation on the integers and then we see that the inverse operation called division is restricted and forces us to introduce the *rational numbers* or fractions. The next step is taking powers, and defining their inverse as taking the corresponding roots. Applied to positive numbers this leads to the *real numbers*, with the remark that of course all rational numbers are real but not the other way around, such as for example the real number $\sqrt{2}$. If we extend the definition of roots to negative numbers we are lead to the introduction of the *complex numbers*, where indeed the fundamental new element is the *imaginary unit* $i = \sqrt{-1}$.

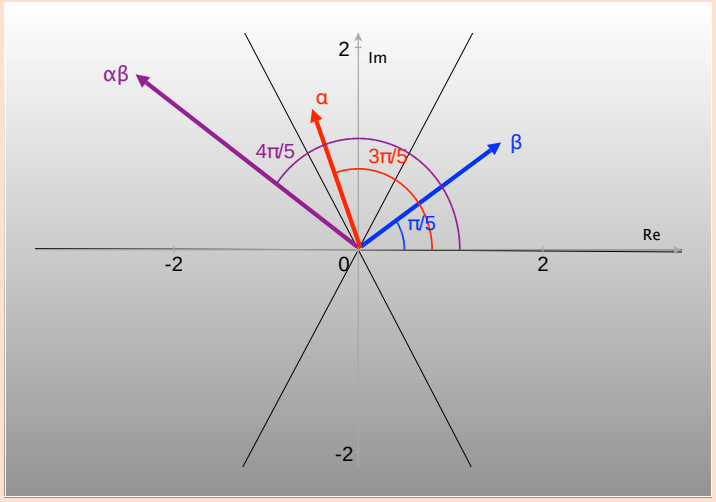
Definition of a complex number. A complex number α has a real and imaginary part $\alpha = a_1 + ia_2$, where a_1 and a_2 are both real, and i is the *imaginary unit* with the defining property $i^2 = -1$. Note that a complex number can therefore also be thought of as a vector in a two-dimensional real space also called the *complex plane*, by taking the real part as the x-component and the imaginary part as the y-component, and thus writing $z = x + iy$. The length of the vector is called the *magnitude or absolute value* of α and denoted by $|\alpha|$, and the angle it makes with the real (x) axis is called its *argument or phase*.
The *complex conjugate* of α is defined as $\alpha^* = a_1 - ia_2$,



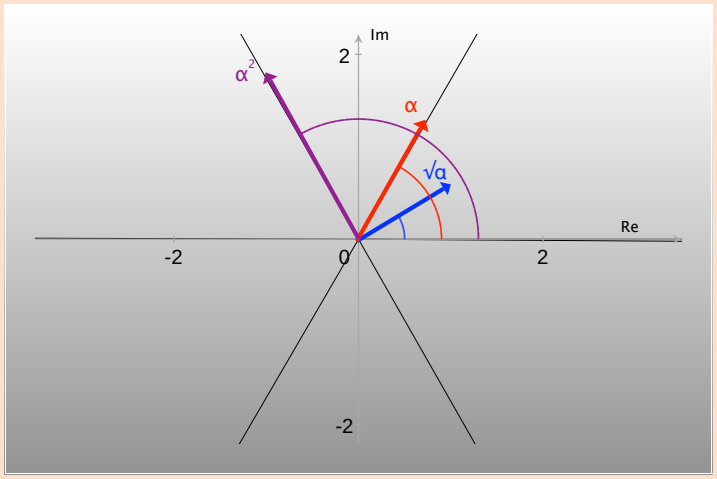
(a) Polar representation of a complex number $\alpha = \rho \exp(i\varphi)$.



(b) Adding and subtracting two complex numbers α and β by the 'parallelogram' rule.



(c) Multiplying two complex numbers α and β amounts to multiplying their magnitudes ($\rho_{\alpha\beta} = \rho_\alpha \rho_\beta$) and adding their phase angles ($\varphi_{\alpha\beta} = \varphi_\alpha + \varphi_\beta$).



(d) The square and square root of a complex number α. Here the blue angle is half, and the purple angle is twice the red angle.

Figure A.20: *Complex numbers.* Graphical representation of some basic operations with complex numbers.

it is obtained by replacing i by $-i$. The value of $|\alpha|$ is defined by the relation $|\alpha|^2 = \alpha^*\alpha = a_1^2 + a_2^2.$, where one obtains the result by multiplying out the expressions and remembering that $-i^2 = +1$, so, $(a_1 + ia_2)(a_1 - ia_2) = a_1^2 - i^2a_2^2 = a_1^2 + a_2^2$. This indeed equals the length of the corresponding vector.

Polar decomposition. There is an alternative but equivalent way to think of complex numbers explicitly using their two-dimensional vector property. If one thinks of a planar vector in polar coordinates, one may specify it by giving its magnitude ρ and the angle φ it makes with the $x-$axis. The complex number is then written as $\alpha = \rho e^{i\varphi}$: the terminology is that φ is called the *argument* or phase angle, and $e^{i\varphi}$ the *phase factor*. We see that $|\alpha| = \rho$ and $|e^{i\varphi}| = 1$. The phase factor describes therefore a point on the unit circle in the complex plane which makes an angle φ with the real axis. This is depicted in Figure A.20(a) from which one also sees that the real part of the phase factor equals $\cos\varphi$, while the imaginary component equals $\sin\varphi$, which leads to a famous mathematical identity originally due to Euler:

$$e^{i\varphi} = \cos\varphi + i\sin\varphi\,. \tag{A.28}$$

This formula is a source of numerous amusing number theoretical identities like $e^{i\pi} + 1 = 0$ and $e^{i\pi/2} = i$. In this parametrization of complex numbers, it is easy to perform complex multiplication and division and taking powers or roots.

Algebraic properties of complex numbers. To add or subtract two complex numbers, one just adds or subtracts their real and imaginary parts separately: $\alpha \pm \beta = (a_1 \pm b_1) + i(a_2 \pm b_2)$. This corresponds to adding (subtracting) two vectors in the plane by the 'parallelogram' rule as indicated in Figure A.20(b). Multiplying two complex numbers α_1 and α_2 amounts top multiplying the magnitudes, i.e. $\rho = \rho_1\rho_2$, while the phase angles add, $\varphi = \varphi_1 + \varphi_2$ as in Figure A.20(c). Similarly when dividing two complex numbers one divides the magnitudes and takes the difference of the phase angles. Taking a complex conjugate amounts to replacing φ by $-\varphi$, i.e. mirroring the vector in the x-axis. We see that the polar representation of complex numbers makes it particularly easy to visualize the multiplication and division operations, but also to take their powers and roots, as we did in Figure A.20(d).♠

♡ On complex vectors and matrices

We have discussed real vectors and matrices in the Math Excursion on page 614. But in quantum theory everything gets complexified, meaning to say that states are represented by complex vectors and observables by complex (hermitian) matrices. Therefore, we will summarize here some additional material specific to complex vectors and matrices.

Complex vectors. Think of our vectors as *column* or *ket* vectors $|v\rangle$ which are complex, which means that the entries or components are complex numbers. Then we may define a space of dual vectors, the dual of a column vector is a *row* or *bra* vector $\langle v|$, *with complex conjugate entries.*

The inner- or dot-product. Having a vector space $\mathcal{V}$ and its dual $\mathcal{V}^*$ the *inner product* between elements of $v^* \in \mathcal{V}^*$ and $w \in \mathcal{V}$ is defined as the number obtained after adding the products of corresponding entries:

$$\langle v|w\rangle = v^* \cdot w = \Sigma_i v_i^* w_i\,.$$

We calculate for example the dot product of two two-dimensional complex vectors as:

$$\begin{pmatrix} 2i & 1 \end{pmatrix} \begin{pmatrix} i \\ 1 \end{pmatrix} = 2i^2 + 1 = -1.$$

The property of the inner product that,

$$\langle w|v\rangle = \langle v|w\rangle^*\,,$$

still implies that $\langle v|v\rangle = \langle v|v\rangle^* = |v|^2$ is always a positive real number which is defined to be the length of the vector $|v|$ squared.

The state space of a qubit. The state of a qubit is by definition the two-dimensional complex vector $|\psi\rangle$ of equation (II.1.2). The normalization condition applied to the state can be written as:

$$\langle\psi|\psi\rangle = |\psi|^2 = |\alpha|^2 + |\beta|^2 = 1\,. \tag{A.29}$$

If we substitute $\alpha = a_1 + ia_2$ and $\beta = b_1 + ib_2$, then we find

$$a_1^2 + a_2^2 + b_1^2 + b_2^2 = 1\,. \tag{A.30}$$

This equation describes a (real) three-dimensional sphere, S^3, embedded in the four-dimensional Euclidean space, R^4, with coordinates (a_1, a_2, b_1, b_2).

Complex matrices acting on complex vectors. Now vectors can also be multiplied by matrices to produce another vector, the way that is done was pictorially indicated for a column vector in A.8(b). This action of matrices on vectors is clearly most easily understood if you think of the matrix as a stack of row vectors. This action can also be considered as a *transformation* of a vector into another vector. A simple example may help:

$$\begin{pmatrix} 1 & i \\ -i & 1 \end{pmatrix}\begin{pmatrix} 2 \\ i \end{pmatrix} = \begin{pmatrix} 2+i^2 \\ -2i+i \end{pmatrix} = \begin{pmatrix} 1 \\ -i \end{pmatrix}.$$

The matrix acts as a *linear operator* on the vector space, as it reshuffles the components into linear combinations of them. Or one may say that $(n\times n)$ matrices map the vector space $\mathcal{V}$ onto itself and we write $A : \mathcal{V} \to \mathcal{V}$. There is for example a particular subset of (3×3) matrices whose action on 'ordinary' vectors corresponds to rotating of those vectors in three-dimensional complex space C^3.

Another example which shows the descriptive power of matrices as operators on state vectors is in (quantum) computation, where generically we think of computation as a sequence of gates, interactions/manipulations or measurements that change the states of a set of (qu)bits.

Such processes or computations can be represented by a product of matrices and rescalings. Indeed the complete computation is just a big operator, mapping the in-state on the out-state vector.

The matrix product. Once we have defined the action of matrices on vectors the step to the multiplication of matrices is straightforward and it was visualized in Figure A.8(c). The (ij)-entry of the product matrix $C = AB$ is obtained by the dot product of the $i-th$ row vectors of A with the $j-th$ column vector of B. Let us again give a simple example:

$$\begin{pmatrix} 1 & i \\ -i & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 1+i & 1-i \\ 1-i & -1-i \end{pmatrix}. \tag{A.31}$$

Types of matrices. As mentioned before, depending on the situation we usually have to put additional constraints defining subsets of matrices, which may or may not be preserved under the basic matrix operations. These definitions involve certain basic matrix manipulations which were represented symbolically in Figure A.10. A fundamental notion is the *transpose* of a matrix denoted by the matrix A^{tr}, which is obtained from A, as we illustrated in Figures A.10(a) and A.10(b). Written in terms of its entries one has $(A^{tr})_{ij} \equiv A_{ji}$. Taking the transpose can therefore also be defined as interchanging rows and columns. Repeating the operation brings you back to the original matrix. Taking the transpose of matrix $C = AB$ we get a matrix which is the product of the transposes, but in the opposite order: $C^{tr} = B^{tr}A^{tr}$. Symmetric or antisymmetric matrices satisfy $A = \pm A^{tr}$ respectively. Note that a symmetric complex matrix contains $n(n+1)$ real numbers, while the antisymmetric one has only $n(n-1)$, it adds up to $2n^2$, the number of real entries of a general complex $(n\times n)$ matrix.

Hermitean matrices. Of special importance in quantum theory are the *hermitian* matrices, because they represent observable physical quantities. To tell you what they look like we first define the hermitian adjoint $A^\dagger$ as $A^\dagger = (A^{tr})^*$ (see Figure A.10(d)). A hermitian (self-adjoint) matrix is just one that satisfies $A = A^\dagger$. It is not hard to see that a hermitian matrix can be decomposed in the sum of a symmetric real and an antisymmetric purely imaginary matrix, also implying that the diagonal elements are real. Such a hermitian matrix contains n^2 real numbers. Let us give a simple example of the above operations for a 2×2 matrix:

$$C = \begin{pmatrix} 1 & i \\ 1 & -1 \end{pmatrix}$$
$$\Rightarrow C^{tr} = \begin{pmatrix} 1 & 1 \\ i & -1 \end{pmatrix}\ ;\ C^\dagger = \begin{pmatrix} 1 & 1 \\ -i & -1 \end{pmatrix}\ ;$$

we see that C is not hermitian because $C \neq C^\dagger$. Each of the Pauli matrices on the left-hand side of equation (A.32) however is hermitian. Note however that their product is *not*.

The Pauli matrices. Most famous are the set of three (2×2) hermitian matrices, which are called the *Pauli matrices* X, Y and Z. They are defined as:

$$X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\ ,\ Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}\ ,\ Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\ , \quad \text{(A.32)}$$

and have a quite unique combination of properties.
(i) They are hermitian: $X^\dagger = X$ etc.
(ii) They are unitary: $X^\dagger X = 1$.
(iii) From (i) and (ii) it follows that they square to the unit matrix: $X^2 = 1$ etc.
(iv) They form a basis of the $su(2)$ Lie algebra, which means that they form a closed algebra under commutation:
$[X, Y] = 2iZ$ etc. (see below).
(v) Their anti-commutator vanishes: $\{X, Y\} = XY+YX = 0$ etc.
(vi) The one qubit observables are linear combinations of the Pauli matrices, the spin-half operators correspond to: $S_x = \hbar X/2$ etc.
(vii) If we add the unit matrix (which commutes will all three of the Pauli matrices, and which is also hermitian), we get the algebra of $u(2) \simeq su(2) \oplus u(1)$.
(viii) Every 2×2 unitary matrix can be written as a linear combination of these four matrices (see below).

Lie algebras.
Hermiticity is not a property that is preserved under matrix multiplication, if you multiply two hermitian matrices their product is not in general. However, their antisymmetric product or commutator is hermitian, so if A and B are hermitian, then:

$$(i[A, B])^\dagger = -i(AB - BA)^\dagger = -i(B^\dagger A^\dagger - A^\dagger B^\dagger) = i[A, B]$$

In this sense the commutator of observables yields another observable, or to put it another way: the observables form a closed commutator algebra, where the 'product' operation of the algebra is then defined as the commutator: $A \cdot B \equiv i[A, B]$. We see a splendid example of this with the qubit where we had three basic observables $\{X, Y, Z\}$ that form a closed algebra under commutation:

$$[X, Y] = 2iZ \quad [Y, Z] = 2iX \quad [Z, X] = 2iY\,, \qquad \text{(A.33)}$$

this three-dimensional algebra is called $su(2)$. The beauty of the subject becomes clear if you think – for example – of the $su(2)$ algebra not as a set of relations that our spin matrices satisfy, but as an abstract set of commutators that define the algebra. In general, one should think of a set of elements X_i that form the basis of the Lie algebra $\mathcal{A}$, satisfying commutation relations:

$$[X_i, X_j] = i \sum_k f_{ijk} X_k\,;$$

the specific set of constants $\{f_{ijk}\}$ are the so-called *structure constants* which define the Lie algebra.

Now you can turn the question around, and ask when given the structure constants, whether there exist any sets of matrices or other operators that actually do satisfy precisely the above relations. This is what one calls the *representation theory* of Lie algebras, an important part of the mathematical theory. In physics we encounter this all the time, for example the $\mathfrak{su}(2)$ algebra is basically the algebra of rotations in three-dimensional space.[3] It is the algebra satisfied by the angular momentum operators $\{L_x, L_y, L_z\}$ as differential operators, but the algebra has also irreducible representation as $(n \times n)$ matrices for any $n = 1, 2, 3, \ldots$. If we write $n = 2s+1$ then s is now defined as the spin, or the angular momentum, and we see that indeed all half-integral and integral values are possible. And the integer values we see recurring as the quantum number l in the spectra of atoms. The $s = 1/2$ case clearly corresponds to the 2×2 matrices S_i . The complex *Lie algebras* and their 'irreducible' representations have been classified completely and form an important subject in the mathematics and physics literature. ♡

[3] This algebra is defined by the commutation relations of equation (A.33) without the factor 2 on the right. In other words $S_x = X/2$ etc.

♢ On symmetry groups

Symmetries are a powerful guiding principle in identifying and understanding important properties of physical systems. The notion of symmetry can be applied to objects, to spaces or lattices, to equations, to the degeneracies in the spectra of atoms and molecules, but also of the electron bands of materials where the ions form an underlying lattice structure. Here we limit ourselves to the basic mathematical background concerning the symmetry groups, which we will refer to throughout the book. In Chapter II.6 we have an extensive section devoted to the physical aspects of symmetries and their breaking.

Groups: the language of symmetry. When we talk about order, we usually refer to some regularities, some predictable pattern that has some or many symmetries. The word symmetry in physics has many different meanings and is like the word 'snow' for the Inuits. One speaks of finite or infinite, discrete or continuous symmetries. Symmetries of objects, of spaces, and of equations. And on another level one speaks of global or local, exact or approximate symmetries. We encountered already the notion of frame rotations, of space-time rotations, and of gauge transformations. And the elaborated structure of fiber bundles as described in chapter I.1, involved the concept of a local or gauge symmetry.

The notions just mentioned are relevant in different contexts but they share the underlying mathematical concept of a *group*. Let us introduce this concept in its elementary easy to grasp form as a *group of transformations*. One can indeed think of transforming an object as applying some operation on it, like rotating it, or moving (translating) it in some direction, or mirroring it (like transforming your left shoe in your right shoe) or scaling the object by changing its size but not its shape. Generally, we think of the group as acting on some vector space, where the objects, like fields or states, are defined as vectors.

Defining properties of a group. Mathematically a group is just a set of elements and a 'product rule' that satisfy some rather obvious axioms, and interestingly those axioms are so restrictive that basically everything is known about the groups that play a role in physics. Group theory is a rich branch of mathematics, and we will only scratch the surface here.

We denote the group by G : it is a set of elements (i.e. transformations or operations) g_i and we write $G = \{g_i\}$ and conversely $g_i \in G$. There are four defining properties:

(i) *composition rule*: if $g_1 \cdot g_2 = g_3$ with $g_1, g_2 \in G$ then $g_3 \in G$, this composition rule is often referred to as the *group multiplication*.

(ii) *associativity*: the group multiplication is associative, which means that the outcome of a product does not depend on the order we perform the multiplication, so,

$$(g_1 \cdot g_2) \cdot g_3 = g_1 \cdot (g_2 \cdot g_3) = g_1 \cdot g_2 \cdot g_3 .$$

(iii) *identity*: there always is the trivial transformation of doing nothing, it corresponds to the identity element e, which satisfies

$$e \cdot g = g \cdot e = g \text{ for all } g .$$

(iv) *inverse*: as you can always transform back, meaning that each element g has a unique inverse g^{-1} with

$$g \cdot g^{-1} = g^{-1} \cdot g = e .$$

Numbers or matrices certainly can form groups but note that we only refer to a single 'composition rule' or 'product' of elements. They do not form a linear space, or an algebra. A set of objects that is closed under some kind of product is maybe the easiest way to think about them. In that sense a group is an elementary and natural notion, and you may be more familiar with it than you think.

Some examples. The set of all integers n form a group

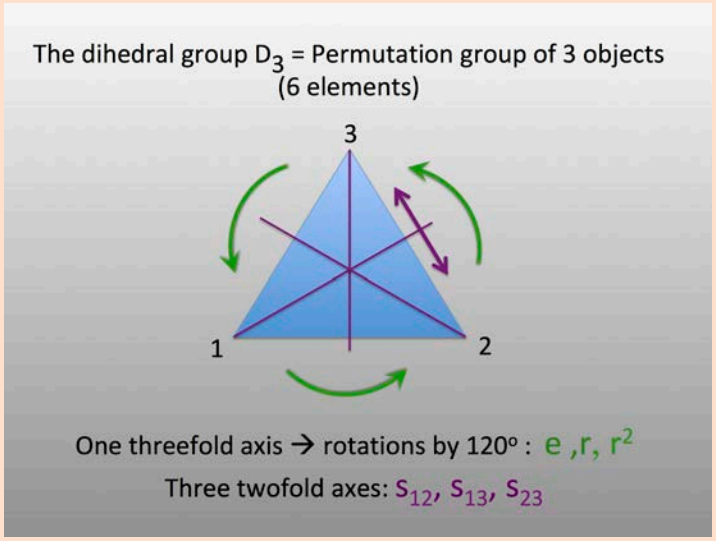


Figure A.21: *The dihedral group* D_3. The symmetry group of an equilateral triangle is the group D_3 consisting of 6 elements. There is one threefold axis, and three twofold axes.

$G = Z$ where the composition rule is addition, the identity element is $n = 0$ and the 'inverse' of n is $-n$. This is an *infinite discrete* group. Note that the integers do *not* form a group under multiplication, because of the problem caused by the inverse operation; zero has no inverse while just dividing two integers brings you outside the integers in to the set of fractional numbers.

The real numbers which correspond to an infinite line form a *continuous* group of translations $T = R$ again under addition (subtraction). Yet another example is by rotations in the plane. We may rotate a two-dimensional object by a certain angle ϕ where $0 \leq \phi < 360^o$. Now the group is not a line but a circle, rotating by 360^o is like doing nothing. This two-dimensional rotation group denoted by $SO(2)$ is the same as the 'phase group', denoted by $U(1)$.

Let us now discuss the group of transformations that leaves some object (or space, or equation) *invariant*, in which case we speak of the *invariance* or *symmetry* group of that object. Consider an equilateral triangle like in Figure A.21;

it is easy to list the transformations that leave it invariant: (i) rotations over 120^o about its center $\{r, r^2\}$, (ii) mirroring it through the bisector of one of the angles $\{s_1, s_2, s_3\}$. This group $G = \{e, r, r^2, s_1, s_2, s_3\}$ has 6 elements and is denoted as the dihedral group D_3. This group is the same as the permutation group S_3 of three objects. The group D_3 readily generalizes for regular polygons (square, pentagon, hexagon...) to groups D_n.

Another important class of groups are groups that leave the inner product of some vector space invariant. For ordinary three-dimensional vectors, the inner product is $a \cdot b = |a||b| \cos\phi$ and the invariance group is the rotation group $SO(3)$. For relativistic four vectors we defined the inner product as $a \cdot b = a_\mu b^\mu = \eta_{\mu\nu} a^\mu b^\nu$, with $\eta_{\mu\nu} = diag(1, -1, -1, -1)$, and it is invariant under the Lorentz group $SO(1, 3)$. In the n-dimensional complex Hilbert space we have state vectors and the hermitian inner product $\langle\Phi|\Psi\rangle$, and as we discussed in this chapter the invariance group is the unitary group $U(n)$. We will have more to say about the unitary groups at the end of this *Math Excursion*.

Space (time) symmetries. In physics and chemistry one type of order refers to the situation where the atoms form a lattice in space and so it is of interest to look at the symmetries of a lattice. If we look at a triangular lattice, or triangular tiling of the plane like in Figure III.2.24(a), we see that we not just have the rotations by multiples of 60^o, but also translations along the sides of the triangles. Those translation can be generated[4] by the two basic translations t_1 and t_2 of the discrete translation group $G = T^2 = T \times T$. Note that each translation group is the same as the group of the integers: $T \simeq Z$.

Abelian versus non-abelian groups. If we now combine the rotations and the translations, we learn something interesting about the structure of the group, namely

[4]Generated means that all translations can be obtained by repeated application of the two basic translations.

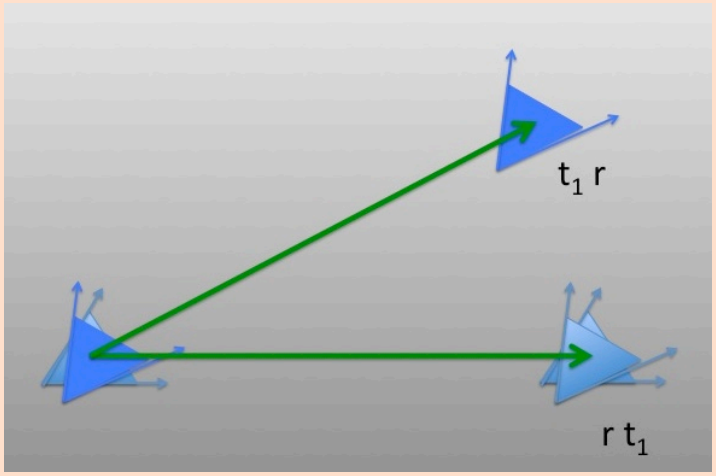


Figure A.22: *The symmetries of two-dimensional Euclidean space.* Picture showing that translations and rotations (of a triangular object) do not commute. It is a fact we are all familiar with: if you make first a step sideways and then turn, you end up in a different place then if you first turn and then make a step sideways. Formally stated: if we first translate along the bottom side of the triangle and then rotate over 30^o, we act with $r \cdot t_1$, and we end up with the rotated triangle in the lower right-hand corner; if we first rotate and then translate, we act with $t_1 \cdot r$, and we end up with the rotated triangle in the upper right-hand corner. The operations are clearly not the same.

that the group composition rule is not necessarily *commutative*, which just means that in general we have that $g_1 \cdot g_2 \neq g_2 \cdot g_1$. The group is then called non-commutative or non-abelian. And this is clearly different from the multiplication and addition of ordinary numbers which are commutative. Ordinary division is of course not, as in general $a/b \neq b/a$, but if you define division as multiplication by the inverse it is, as $a\frac{1}{b} = \frac{1}{b}a$.

The rotations (in a plane) by themselves do commute, if I first rotate by an angle ϕ_1 and then ϕ_2 the net result is a rotation by $\phi_1 + \phi_2$, and that is the same as first rotating by ϕ_2 and then by ϕ_1. The same is true for the translations by themselves as $a + b = b + a$.

It is no longer true if we combine rotations and translations as we did in in Figure A.22. If we choose r with $\phi = 60^o$ and $t_1 = na$ translation over n times side of a triangle, then both operations leave the lattice of Figure III.2.24(a) invariant. They belong to the invariance group of the lattice but correspond to different elements. The terminology is that we call the total invariance group of a lattice a *space group* whereas the rotational part of it forms a *point group* as it leaves a point of the space fixed. Note that if we think of the plane as a continuous space, usually denoted by R^2, then the space group would be the group made up by arbitrary rotations and arbitrary translations; this is a continuous group denoted by E_2, the Euclidean group in two dimensions. Also this group has of course higher n-dimensional analogues called E_n.

Groups of matrices. There are many groups that can be represented by matrices, because square matrices close under the matrix product. Generically such groups are non-abelian. But one can also make restrictions to subsets of matrices that form closed subsets under matrix multiplication. Of special interest for us are the *orthogonal* and *unitary* matrices $O(n)$ and $U(n)$. They act as non-abelian transformation groups of rotations on the real and complex spaces $\mathbb{R}^n$ and $\mathbb{C}^n$. The matrices satisfy $O\ O^{tr} = 1$ and $U\ U^\dagger = 1$ respectively.

The group $SU(2)$ of 2×2 unitary matrices. 
Let us add an important remark on the relation between hermitian and unitary matrices. Let me recall the Euler formula for the exponential of imaginary number '$i\varphi$' (A.28):

$$e^{i\varphi} = \cos\varphi + i\sin\varphi\,.$$

The sine and cosine appearing show that it is indeed a periodic function, and therefore, we choose an angular variable φ. You might wonder whether similar formulas can be written down for matrices. The answer is a full-fledged yes, and that brings us to the relation between Lie algebras and Lie groups. Let me give you the extremely useful generalization of the Euler formula to the Hermitean (2×2) matrices. Consider an $su(2)$ matrix,[5]

$$A = (\hat{n}_x X/2 + \hat{n}_y Y/2 + \hat{n}_z Z/2)\,,$$

where $\hat{n}$ is some arbitrary vector of unit length and θ some angular variable, then in general, the following relation holds:

$$e^{i\theta A} = \mathbf{1}\cos\theta/2 + iA\sin\theta/2\,. \qquad \text{(A.34)}$$

This elegant equation has many applications in all venues of theoretical physics, and we will use it repeatedly later on. It does for example represent a rotation of a two-component spinor over an angle θ around the $\hat{n}$ axis, with the peculiar but characteristic property that a rotation by $\theta = 2\pi$ of any spinor maps it to minus itself. As mentioned before, that is a property that distinguishes spinors from 'ordinary' vectors. One thing that is immediately clear from the above formula, is that the expression corresponds to a unitary matrix. This holds in general: if we write a matrix U as an exponential of a hermitian matrix A, then we can write:

$$U^\dagger = (e^{iA})^\dagger = e^{(iA)^\dagger} = e^{-iA^\dagger} = e^{-iA} = U^{-1}\,, \qquad \text{(A.35)}$$

which shows that U is a unitary matrix. This property that exponentials of hermitian matrices are unitary operators is widely used in quantum theory, in particular in the theory of (unitary) representations of symmetry groups that act on the Hilbert space of a system.

So to summarize this part we saw a close relationship between the 'algebra of observables' for a quantum system, being a Lie algebra, i.e. a closed commutator algebra, which when put in the exponent yields a corresponding Lie group. In that sense we say that the observables (the Lie algebra) generate small or infinitesimal transformations, while the exponents (elements of the Lie group) correspond to finite transformations.

[5] We have mentioned before that half the Pauli matrices $\{X/2, Y/2, Z/2\}$ do form a basis for the angular momentum or spin algebra, as they satisfy $[S_x, S_y] = iS_z$ etc.

Invariants. There are two more properties of matrices we want to discuss: these are what are called invariants under basis transformations. First observe that we may rotate the basis of a vector space. Then the components of the vector change and are obtained by acting with the corresponding matrix U. In the main text we showed that basis transformations must preserve the scalar product of two arbitrary vectors, and therefore will satisfy the unitarity condition $U^\dagger U = 1$, and therefore $U^\dagger = U^{-1}$. So if we have a matrix operator A acting on vectors in a given frame and we ask what the matrix looks like in the rotated or 'primed' frame we can see that from the following algebraic manipulations. First we define:

$$|\psi'\rangle \equiv U\,|\psi\rangle \text{ and } |\phi\rangle \equiv A|\psi\rangle\,,$$

which allows us to write:

$$\begin{aligned} |\phi'\rangle &\equiv U\,|\phi\rangle = U\,A|\psi\rangle \\ &= UAU^{-1}U|\psi\rangle = UAU^{-1}|\psi'\rangle = A'|\psi'\rangle \end{aligned}$$

Implying that $A' = UAU^{-1}$. Given these expressions for how state vectors and observables transform under unitary basis transformations, you might ask whether there are any quantities related to these observables that are preserved under such transformations. The answer is affirmative: the invariant quantity corresponds to the set of eigenvalues, particularly the sum and the product of all eigenvalues, denoted as the *trace* and the *determinant*.

The *trace* of a matrix A denoted by tr A is defined as the sum of the diagonal elements, so tr $A = \Sigma_i A_{ii}$. The trace is indeed invariant under basis transformations as one easily sees:

$$\text{tr } A' = \text{tr } (UAU^{-1}) = \text{tr } (U^{-1}UA) = \text{tr } A\,.$$

The trace satisfies the cyclic property meaning that the trace of a product is invariant under cyclic permutations, i.e. that is putting the matrices in the trace on a circle holding hands and moving them around:

$$\begin{aligned} \text{tr } (ABC) &= \Sigma_{ijk}(A_{ij}B_{jk}C_{ki}) = \\ &= \Sigma_{ijk}(C_{ki}A_{ij}B_{jk}) = \text{tr } (CAB) \text{ etc.} \end{aligned}$$

The point is that all indices are pairwise summed over. We will see that the trace, because it is frame independent, plays an important role in certain aspects of quantum theory. ◊


1994
CUBA
CORREOS
MAX PLANCK
1858 - 1947
40
CELEBRIDADES DE LA CIENCIA
h=6,626·10⁻³⁴Js
E=h·ν
55
Deutschland
1858-1947
UNION DES COMORES
225 FC.
2009
DECOUVERTES CELEBRES
DEMOKRATISCHE
DEUTSCHE
REPUBLIK
MAX PLANCK 10 1858-1947
DEMOKRATISCHE
DEUTSCHE
REPUBLIK
MAX PLANCK 20 1858-1947
30
MAX
PLANCK
DEUTSCHE POST
BERLIN
COSTA RICA CORREOS
ħ = 6.626 x10⁻³⁴ Js
Max
Planck
95
colones
Año Mundial de la Física
N$ 0.60
AEREO
MAX PLANCK
1858-1947
PREMIO NOBEL
DE FISICA 1918
75 AÑOS
PREMIO NOBEL
URUGUAY
GHANA ¢400
GERMANY
MAX PLANCK
(1858-1947) PHYSICS 1918
NOBELPRIS 1918
170
SVERIGE
PLANCK
PRIX NOBEL DE PHYSIQUE 1914
LUDWIG ERNST KARL MAX PLANCK
500F
REPUBLIQUE DE CÔTE D'IVOIRE
REPUBLIQUE GABONAISE
500f
M. PLANCK
Postes
2008

Appendix B

Chronologies, ideas and people

In this appendix we list the scientific achievements in the quantum domain over more than a century as well as the names and the dates of the Nobel prizes that were awarded for these. It demonstrates the fact that quantum is everywhere and overtook progress in physics to a large extent.

The tables cover the following topics:

Figure B.1: The early quantum giants at the fifth Solvay conference, held in Brussels in 1927. On that occasion quantum mechanics, including the ‘Copenhagen interpretation’, was presented as a complete and final theory of atomic phenomena.

The person	Year	The concept	The mathematical statement
Planck	1897	Planck's constant	$\hbar = h/2\pi$
	1900	Black-body radiation	$\rho(\nu, T) = \frac{8\pi V \nu^2}{c^3} \frac{h\nu}{(e^{h\nu/kT} - 1)}$
Einstein	1905	Photoelectric effect, the photon	$E = h\nu$
Bohr	1913	Atomic model	$E_n \sim \hbar^2 e^2 / 2mc^2 n^2$
De Broglie	1923	Matter waves	$\lambda = \hbar / mv$
Einstein, Podolski, Rosen	1920	EPR paradox, entanglement	$\|\psi(1,2)\rangle = (\|00\rangle \pm \|11\rangle)/\sqrt{2}$
Bose, Einstein	1924	Quantum statistics, Bose condensate	$n_i = g_i/(e^{\beta(\epsilon_i - \mu)} - 1)$
Pauli	1924	Exclusion principle	$\psi(x_1, x_2) = -\psi(x_2, x_1)$
Heisenberg	1925	Matrix mechanics	$d\hat{A}/dt = i[\hat{H}, \hat{A}]$
	1927	Uncertainty relations	$\Delta x \, \Delta p \geq \hbar/2$
Von Neuman	1925	Density matrix, quantum entropy	$\rho = \Sigma p_a \|\psi_a\rangle\langle\psi_a\|, \; S = tr(\rho \ln \rho)$
Schrödinger	1926	Wave mechanics	$i\hbar \, d\psi/dt = \hat{H} \, \psi$
Born	1926	Probability interpretation	$\psi = \Sigma \, c_i \, \chi_i \Rightarrow P_{\lambda_i} = \|c_i\|^2$
Fermi	1927	Quantum statistics for fermions	$n_i = g_i/(e^{\beta(\epsilon_i - \mu)} + 1)$
Dirac	1927	Dirac equation	$(i\hbar\partial\!\!\!/ + eA\!\!\!/ + m)\,\psi(x,t) = 0$
Bell	1964	Bell inequality	$\|P_c(a,b) - P_c(a,c)\| \leq 1 + P_c(b,c)$
Bennett, Brassard, Deutsch, Shor	>1980	Quantum information/computation	key distribution, teleportation, prime factoring algorithm

Table B.1: Foundational quantum concepts and their protagonists.

Kamerling Onnes	Superconductivity (experiment)	1911
Bloch	Conduction band	1920
Uhlenbeck, Goudsmit	Spin	1925
Van Vleck	Theory of magnetism	1935
Kapitza, Allen, Misener	Superfluidity	1938
Pauling	The nature of chemical binding	1939
Rabi	Nuclear magnetic resonance (NMR)	1946
Purcell, Bloch	NMR (implementations)	1952
Bardeen, Houser, Brattain, Shockley	Semiconductors, Transistor	1950
Gabor	Holography	1950
Landau	Fermiliquids, quasiparticles, phase transistions	1952
Bardeen, Cooper, Schrieffer	BCS theory of superconductivity	1957
Townes, Basov, Prokhorov	Laser	1958
Anderson	Localization	1958
Ahoronov, Bohm	Aharonov-Bohm effect	1959
Haldane, Kosterlitz, Thouless	Topological phase transitions	1973
De Gennes	Liquid crystals (mostly classical physics)	1974
Laughlin	Theory of Fractional Quantum Hall effect	1983
Berry	Berry phase	1984
Cornell, Wiegmann	Bose Einstein condensation (experiment)	1995
Kitaev, Wen	Topological order	1997
Lauterbur, Mansfield	Magnetic resonance imaging (MRI)	2003
Geim, Novoselov	Graphene	2004
Aspect, Clauser, Zeilinger	Entangled photons (experiments)	>1980

Table B.2: Turning points in quantum condensed matter (theory) and quantum optics.

Feynman, Swinger, Dyson, Tomonaga	Quantum electrodynamics (QED)	1946
Yang, Mills	Non-Abelian gauge theory	1954
Gellmann, Zweig	SU(3) Quarks	1963
Nambu, Jona Lasinio	Chiral symmetry breaking	1965
Glashow, Weinberg, Salam	Weak and electromagnetic theory	1968
Higgs, Brout, Englert	Higgs mechanism	1969
't Hooft, Veltman	Renormalization of non-Abelian gauge theories	1970
Wilson	Theory of critical phenomena, confinement	1972
Gellmann, Leutwyler, Fritsch	Quantum Chromodynamics (QCD)	1971
Gross, Politzer, Wilczek	Asymptotic freedom	1973
Witten, Schwarz, Green	String theory	1983
Polyakov, Belavin, Zamolodchikov	Conformal Field Theory (CFT)	1983
Witten	Topological Field Theory	1983
Maldacena	Anti de Sitter/CFT correspondence	1995

Table B.3: Turning points in Elementary particle theory.

Röntgen	X-rays	1901
Becquerel, Curie, Curie	Radioactive decay (α and β radiation)	1903
Thomson	Electron	1906
Rutherford	Nucleus	1908
Planck	Quanta of radiation	1918
Einstein	Photon	1921
Compton	Compton effect	1927
Chadwick	Neutron	1935
Anderson	Positron	1936
Powell	Pion	1950
Chamberlain, Segre	Antiproton	1959
Richter, Ting	J/Psi meson	1976
Rubia, Van der Meer	W and Z bosons	1984
Lederman, Schwartz, Steinberger	Muon neutrino	1988
Friedman, Kendall, Taylor	Quarks	1990
Perl	Tau-neutrino	1995
Reines	Neutrino	1995

Table B.4: Nobel prizes awarded for discovery of elementary particles.

Bethe	Energy production in stars	1967
Ryle, Hewish	Pulsars	1974
Penzias, Wilson	Microwave background radiation	1965
Chandrasekhar, Fowler	Theories of star evolution	1983
Hulse, Taylor	Precision tests of gravity	1993
Davis, Koshiba	Cosmic neutrino's	2002
and Giacconi	X-ray sources	
Mather, Smoot	Anisotropy in background radiation	2006
Perlmutter, Schmidt, Riess	Accelerated expansion	2011
Thorn, Weiss, Barish	Gravitational wave detection	2017

Table B-7: Nobel prizes for astrophysics and cosmology.

Rabi	Nuclear magnetic resonance	1944
Bridgman	Apparatus to produce extremely high pressures	1946
Blackett	The Wilson cloud chamber method	1948
Powell	Photographic method of studying nuclear processes	1950
Bloch and Purcell	Nuclear magnetic precision measurements	1952
Zernike	Phase contrast microscope	1953
Glaser	Bubble chamber	1960
Shockley, Bardeen and Brattain	Transistor	1956
Alvarez	Hydrogen bubble chamber and data analysis techniques	1968
Gabor	Holographic method	1971
Ryle and Hewish	Radio astrophysics	1974
Bloembergen and Schawlow	Laser spectroscopy	1981
Siegbahn	High-resolution electron spectroscopy	1981
Ruska	Electron microscope	1986
Binnig and Rohrer	Scanning tunneling microscope	1986
Ramsey	Separated oscillatory fields method and its use in atomic clocks	1989
Dehmelt and Paul	Ion trap technique	1989
Charpak	Multiwire proportional chamber	1990
Brockhouse	Neutron spectroscopy	1994
Shull	Neutron diffraction	1994
Alferov and Kroemer	Semiconductor heterostructures, high-speed- and opto-electronics	2000
Kilby	his part in the invention of the integrated circuit	2000
Hall and Hänsch	Laser-based precision spectroscopy, optical frequency comb technique	2004
Kao	Light transmission in fibers for optical communication	2009
Boyle, Smith	invention of imaging semiconductor circuit - the CCD sensor	2009
Fert and Grünberg	Giant magnetoresistance	2007
Haroche, Wineland	Measuring and manipulation of individual quantum systems	2012
Akasaki, Amano, Nakamura	Bright blue light-emitting diodes	2014
Weiss, Barish, Thorne	Gravitational wave detector LIGO	2017

Table B.6: Nobel prizes awarded (from 1944 onwards) for the invention and development of new techniques and devices.

Indices

Subject index Volume III

Name index Volume III

List of Figures

List of Tables

Recommendations

‘ This beautiful three-part journey is an insightful and profound exploration of the hidden depths of the quantum realm and its far-reaching implications and applications. Beyond your typical popular science book, The Power of the Invisible offers readers an immersive and enlightening experience that goes far beyond the surface level. With Bais as your guide, you?ll be granted a rare glimpse into the unseen world that lies beneath the fabric of our reality.’

Harry Buhrman
Professor of Quantum Computation, University of Amsterdam, Director QuSoft, Chief Scientist *Quantinuum*

‘ Power of the Invisible focuses on the explosion of quantum mechanics into our world, from its origins in the early 20th century to its manifestation as computing technology in the 21st, with clear, and often mathematical, explanations of its myriad consequences and extrapolations, from liquid Helium to the Standard model, to superstrings... his account of this domain makes a subject with a difficult reputation wonderfully transparent.’

Michael Freedman
Fields medal recipient, Professor of mathematics, UC Santa Barbara, Founding Director Station Q Microsoft

‘ This remarkable book is a tribute to the magic of science. [...] It is a bible for aspiring physicists to catapult their worldview and scientific maturity light years ahead; a compendium for older scientists to reflect on the world and fill holes in their knowledge; and a gorgeous coffee table book for any aficionado of science who wants to contemplate the wonders of the universe. Most of all, it is fun to read and ready to become a cult classic.’

J. Doyne Farmer
Baillie Gifford Professor of Complex Systems Science, Smith School of Geography and Environment, Oxford University, Founder of Prediction Company, and Chief Scientists of Macrocosm Ltd.

‘In three masterfully written and beautifully illustrated books Bais gives he interested reader a unique and comprehensive overview of the foundations, the inner workings, and the far-reaching implications of the quantum revolution.’

Erik and Herman Verlinde
Professors of Theoretical Physics, University of Amsterdam and Princeton University

Acknowledgements

The creation of the book has been an exciting journey and I am indebted to many people who have helped me along the way. These include my colleagues from the Institute for Theoretical Physics in Amsterdam, of whom I like to mention Jan Smit, Chris van Weert, Karel Gaemers, Leendert Suttorp, and Jan Pieter van der Schaar, as well as a number of former students. Indeed: *teaching is the ultimate way of learning*. I am grateful to my former teachers and advisers: Profs Hans van Leeuwen, Michael Nauenberg, Joel Primack, and indirectly Gerard 't Hooft. I am indebted to Erik and Herman Verlinde, Robbert Dijkgraaf, Kareljan Schoutens and Harry Buhrman for thoroughly enriching my quantum perspectives. My views on science in general have been deeply influenced by my collaborators and colleagues at the Santa Fe Institute, in particular with respect to the fundamental notions of computation and evolution.

I thank Dr Manus Visser for a superb job on a thorough and critical proofreading the whole work and suggesting very many ways to improve it. I am indebted to Jan Peter Wissink, director of AUP, for his patience and persistent support during the long journey towards completion of this work. I thank editor Evelien Witte–Van der Veer for coordinating the production at AUP. I have profited from conversations and advice from Peter Ghijsen and Lucy Wenting on matters of style and layout, and last but not least with Doyne Farmer on science in general and the art of writing semi-popular science books.

I lam indebted to the organizations, institutes and individuals that have through their financial support made the publication of this quantum trilogy possible. And I thank Sijbolt Noorda and Joost van Mameren for initiating and coordinating the fundraising effort.

Last but not least, I like to thank my wife Vera, and my beloved children for their continued encouragement and warm support during this mission.

About the Author:

Sander Bais studied applied physics at Delft University in The Netherlands and obtained a PhD in Theoretical Particle Physics from UCSC and SLAC in the US in 1978. He was a research fellow at the University of Pennsylvania, and scientific associate at CERN, and became full professor of theoretical physics at the University of Amsterdam in 1985. He was associated with the Santa Fe Institute as external professor from 2007 until 2020.

Sander Bais has been director of the Institute for Theoretical Physics of the University of Amsterdam, member of the governing board of the NWO/FOM funding agency, and scientific delegate in the CERN Council.

His active research focused on topological aspects of gauge and string theory with applications to both particle and condensed matter physics, He also made regular excursions to astrophysics.

He is the author of number of successful semi-popular books on theoretical physics that have been translated in more than 15 languages.

Power of the Invisible

Quantum Physics is the solid basis of most of our understanding of nature and has been the driver of many technological advances. The trilogy *Power of the Invisible: The Quintessence of Reality* gives a coherent account of this huge domain of knowledge, which is linked to some fifty Nobel prizes and is one of the greatest scientific achievements of the twentieth century. This quantum story follows three lines in parallel: a pictorial, an explanatory and a mathematical one.

Sander Bais is a distinguished theoretical physicist now retired from the University of Amsterdam; previously, he was a long-time external faculty member of the Santa Fe Institute. He has made pioneering contributions to topological physics with applications varying from high-energy to condensed matter physics. His popular books, including *Very Special Relativity*, *The Equations* and *In Praise of Science* have been published in more than 15 languages.

"In three masterfully written and beautifully illustrated books Bais gives he interested reader a unique and comprehensive overview of the foundations, the inner workings, and the far-reaching implications of the quantum revolution."
Erik and Herman Verlinde, Professors of Theoretical Physics, University of Amsterdam and Princeton University

Amsterdam University Press

9 789048 562879