

THE FRONTIERS COLLECTION

Wouter Schmitz

PARTICLES, FIELDS AND FORCES

A Conceptual Guide to Quantum Field Theory and the Standard Model

Second Edition



 Springer

THE FRONTIERS COLLECTION

Series Editors

Avshalom C. Elitzur, Iyar, Israel Institute of Advanced Research, Rehovot, Israel

Zeeya Merali, Foundational Questions Institute, Decatur, GA, USA

Maximilian Schlosshauer, Department of Physics, University of Portland, Portland, OR, USA

Mark P. Silverman, Department of Physics, Trinity College, Hartford, CT, USA

Jack A. Tuszyński, Department of Physics, University of Alberta, Edmonton, AB, Canada

Rüdiger Vaas, Redaktion Astronomie, Physik, bild der wissenschaft,
Leinfelden-Echterdingen, Germany

The books in this collection are devoted to challenging and open problems at the forefront of modern science and scholarship, including related philosophical debates. In contrast to typical research monographs, however, they strive to present their topics in a manner accessible also to scientifically literate non-specialists wishing to gain insight into the deeper implications and fascinating questions involved. Taken as a whole, the series reflects the need for a fundamental and interdisciplinary approach to modern science and research. Furthermore, it is intended to encourage active academics in all fields to ponder over important and perhaps controversial issues beyond their own speciality. Extending from quantum physics and relativity to entropy, consciousness, language and complex systems—the Frontiers Collection will inspire readers to push back the frontiers of their own knowledge.

Wouter Schmitz

Particles, Fields and Forces

A Conceptual Guide to Quantum Field
Theory and the Standard Model

Second Edition



Springer

Wouter Schmitz
Amsterdam, Noord-Holland
The Netherlands

ISSN 1612-3018 ISSN 2197-6619 (electronic)
THE FRONTIERS COLLECTION
ISBN 978-3-030-98752-7 ISBN 978-3-030-98753-4 (eBook)
<https://doi.org/10.1007/978-3-030-98753-4>

1st edition: © Springer Nature Switzerland AG 2019
2nd edition: © The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer
Nature Switzerland AG 2022

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

*To Ingeborg,
Eline, Ewout and Leonore
With love*

Contents

1	Introduction	1
2	Particles or Waves?	3
2.1	How to Describe a Wave	6
2.1.1	Wavelength Represents Momentum	6
2.1.2	Frequency Represents Energy	8
2.1.3	Superposition and Interference of Waves	10
2.1.4	Measurement	12
2.2	Probability Amplitude	14
2.3	What Is Waving?	16
3	Fields and Waves Making Up Reality	19
3.1	What Is a Field?	19
3.2	All We Are Is Waves in a Field	21
3.2.1	Objection 1: How Can a Billiard Ball Be a Wave?	21
3.2.2	Objection 2: But Things Do Not Look Like Waves	22
3.2.3	Objection 3: How Can Waves Make a Table Seem Massive?	22
3.2.4	Objection 4: But Waves Die Out, Don't They?	23
3.2.5	Objection 5: What, Then, Is Empty Space, Through Which I Can Throw a Ball?	23
3.2.6	Objection 6: Oh-No, Not the Ether Again...	24
3.3	Conclusion	26
4	What Is a Particle If It Is a Wave?	29
4.1	Where Is a Particle?	29
4.2	Waves in Space	30
4.3	Waves in Space and the Double Slit Experiment	34
4.4	Waves in Time	36
4.5	A Particle Is a Bunch of Waves	37
4.6	Velocity of Particles and Waves	38

5	The Potential of a Field's Elasticity	41
5.1	Exchanging Energy in a Field	41
5.2	Waves in a Medium	44
6	A Wave of Relativity	49
6.1	Wave Velocity	49
6.2	How Does a Wave Become Massive?	50
6.2.1	A Game with Rope and Springs	50
6.2.2	Consequence 1: You Cannot Go Faster Than Light	52
6.2.3	Consequence 2: The Relation Between Frequency and Wavelength Depends on the Mass	52
6.2.4	Consequence 3: Mass = Inertia	56
6.2.5	Consequence 4: Other Potential Differences Can Create "Mass"	59
6.2.6	Consequence 5: Mass Can Be Changed Into Energy	61
6.2.7	Example: Photons in a Plasma	61
6.2.8	Conclusion and Summary	62
6.3	The Elasticity of the Minkowski Metric	63
6.4	Length Contraction and Time Dilation of Waves	66
6.5	About Higgs	69
7	Quantization of Fields	71
7.1	First Quantization	71
7.2	Second Quantization	73
7.3	Phonons	76
7.4	Conclusion	77
8	Energy in Waves and Fields	79
8.1	Conservation Laws	80
8.1.1	Energy—Momentum Tensor	82
8.2	How to Envision a Field Quantum	85
8.2.1	Coupled Oscillators	85
8.3	Annihilation of a Field Quantum	89
8.4	Describing a Field Quantum	90
9	Symmetry and the Origin of Force	93
9.1	Rotational Symmetry	93
9.1.1	Rotations in a Plane	93
9.1.2	Rotations in Three Dimensions	95
9.1.3	Rotations in Eight Dimensions	97
9.2	The Electromagnetic Field	98
9.2.1	QED	103
9.2.2	The Electromagnetic Field	103
9.3	Path Integral	106
9.4	How Does Symmetry Create a Force?	110
9.5	A Constant Field and the Refractive Index	116
9.6	Conclusion Regarding the Electromagnetic Force	118

10 Propagators and Virtual Particles	119
10.1 Time Order of Events and Feynman Diagrams	123
10.2 Propagator	129
10.3 Relation Between Virtual and Real Particles	134
10.3.1 Summarizing	135
10.4 What Is an Electron Really?	135
10.5 How Do Virtual Particles Create a Force?	138
10.5.1 Electrons of Equal Charge	138
10.5.2 An Electron and a Positron	141
10.5.3 Conclusion	143
10.6 Path Integral Revisited	144
10.7 Fluctuating Fields	146
10.7.1 Casimir effect	146
10.8 The Arrow of Time	147
11 Renormalisation of Fearful Infinities	153
11.1 Renormalizing Mass	155
11.2 Renormalizing Charge	156
11.3 Renormalisation Group	158
12 Is the Cat Dead or Alive? How Quantum Decoherence ‘Digitized’ the Universe	161
12.1 Quantum De-Coherence and “Collapse of the Wave Function”	162
12.1.1 Entanglement	163
12.1.2 Quantum Decoherence	164
12.1.3 The Transition from Quantum Behaviour to Classical Behaviour	166
12.2 Tunnelling and Decoherence	170
12.3 Quantum Computing	173
12.4 Schrödinger’s Cat	178
13 Spin Makes Up Bosons and Fermions	179
13.1 What Is Spin?	179
13.1.1 Orbital Momentum in the Atom	180
13.1.2 The Origin of Spin	185
13.1.3 Spin as a Wave	189
13.2 Fermions and Bosons	192
13.2.1 Wave Phase	193
13.2.2 Fermions	193
13.2.3 Bosons	195
13.2.4 Spinor Fields	196
13.2.5 Fermions in Opposite Spin	200
13.3 Helicity	201
13.4 Chirality	202

13.4.1	Fermions Come with Two Chiralities, Called Left and Right. Bosons Do Not	202
13.4.2	Under Parity, the Chirality of a Fermion Is Swapped to the Opposite Chirality	203
13.4.3	Low Velocity Fermions Flip Chirality at the Frequency of Their Mass	205
13.4.4	Chirality Is Not the Same as Spin	206
13.4.5	Fermions of Different Chirality Are Different Particles	206
13.4.6	At Very High Velocities, the Chirality of Fermions Becomes Fixed and Related to Their Helicity	207
13.5	Fermions Becoming Bosons	210
13.6	Conclusions on Spin, Helicity, and Chirality	212
14	Conservation of Charge and Particle Number	213
14.1	Particle Number Conservation	213
14.2	Charge Conservation	214
15	Particle Zoo	217
15.1	A Visit to the Particle Zoo	217
15.2	Introducing the Fundamental Particle Overview	224
15.3	The Rabbit Hole	225
16	Electroweak Force in the Early Universe	229
16.1	The First 10^{-12} s	230
16.1.1	Electron and Neutrino Waves	231
16.1.2	Introducing the Original U(1) Gauge Field	232
16.2	Symmetry Amongst the Waves	233
16.2.1	Introducing the SU(2) Gauge Field	235
16.2.2	Including Isospin Symmetry in the Overview of Waves	239
16.3	Introducing the Higgs Field	240
16.3.1	Fields Overview First 10^{-12} s	243
16.4	Fundamental Particle Overview 2	244
17	Symmetry Breaking and the World Was Never the Same Again	247
17.1	Mixing Fields	248
17.1.1	What Condensates in the Vacuum?	249
17.2	Breaking the Symmetry of the Higgs Field	251
17.2.1	Consequences for the U(1) and SU(2) Gauge Bosons	252
17.2.2	Mass of the W^- , W^+ and Z°	256
17.2.3	Consequences for the Fermion Interaction Potentials	257
17.3	Interactions	259
17.3.1	Photon Interactions	259
17.3.2	W-Interactions	261

17.3.3	Z-Interactions	264
17.3.4	Z-W Self-interactions	265
17.3.5	Neutron Decay	266
17.3.6	Radioactive Decay	267
17.3.7	Cabibbo Rotation	269
17.3.8	Neutrino Oscillations	271
17.3.9	Concluding	274
17.4	Fermions Gaining Mass	274
17.5	Parity Violation and CPT Symmetry	276
17.6	Family Business	282
17.7	Fundamental Particle Overview 3	284
18	The Strong Force: Quantum Chromodynamics	287
18.1	The Big Why	287
18.2	The Colour Symmetry	290
18.3	QCD Fields: Overview	299
18.3.1	Colour Confinement	300
18.3.2	Quark Jets	303
18.3.3	Asymptotic Freedom	305
18.4	Composite Particles	307
18.5	Interactions	309
18.5.1	Quark–Quark Colour Interactions	310
18.5.2	Annihilation and Creation	312
18.5.3	Gluon–Gluon Interactions	313
18.5.4	Proton–Anti-proton Collisions	313
18.5.5	Residual Strong Force or “Nuclear Force”	314
18.6	Masses of Quarks, Mesons, and Baryons	317
18.7	Fundamental Particle Overview 4	318
19	Gravity as a Field	321
19.1	A Field Theory of Gravity	321
19.2	Background Independence	326
19.3	Other Problems	327
20	Further Reading	331
20.1	Pop Science	331
20.2	The Internet	332
References and Sources	335	
Index	343	

Chapter 1

Introduction



In the past century science has made a lot of progress in understanding the world of fundamental particles and forces. This fundament of the world we live in has always fascinated me enormously. And I am not alone in this. Unfortunately, it turns out that these building blocks of the universe are as hard to understand as they are interesting. This mix makes it a magic world of exotic particles and weird forces that are the realm of the modern magicians we call physicists. In ancient times magicians used spells that no-one could understand. Our modern magicians use mathematics as their language to get a grip on the universe. Of course, where ancient magicians failed to understand the world and we have learned that their spells do not work, this is entirely different for present-day physicists.

Physics is successful in understanding the universe since calculations agree with measurements to sometimes very high precision. So, physicists say that they understand the systematics of our world at least to some level. Mathematics plays a crucial role in these calculations. Without very advanced mathematics, the theories that physicists have created would be impossible to describe. Often, new theories could be developed only after mathematics had laid the groundwork with which to describe them.

Personally, I found this very frustrating. We can calculate some things in the universe, but how should I imagine it? To me, understanding is not about being able to calculate it. I would like to understand “how things work” by being able to imagine them. Understanding them in words and pictures rather than in formulas.

So, I went to study physics, hoping to become one of the magicians. But I ran into trouble: in those days (we are talking around 1990) most physicists were very much about calculating (and probably they still are). One of them said to me: “if you can calculate it, you understand it” leaving me in despair.

Again unfortunately, I have to agree with them. Any time you try to get a picture of how something works, you easily go wrong. It is very easy to “reason” towards a result that just isn’t true. All too often, it is only by using mathematics that the logic prevents you from taking the wrong turn.

But the emptiness remained. So, I kept on working on understanding what all that mathematics actually means. Well aware of the dangers of using metaphors, pictures, and words I still went ahead in my search for the holy grail: to get some grasp of what is going on in these theories.

The high point of all the well tested and fundamental theories is quantum field theory and the standard model, which is expressed in terms of quantum fields. I worked through the mathematics, but also read authors that came up with metaphors. Then I tried to match metaphors to the mathematics. One thing I discovered is that there are many metaphors out there that seem easy to understand but are simply too far from the mathematics to be taken seriously.

I also worked the other way: trying to get a picture of what a mathematical formula means, could I find one that would work according to the same mathematics so that it might serve as a good metaphor? After a long but serious effort I came up with the book you have in front of you. It contains a story of how it all works in terms of metaphors, pictures, and words and I have tried to be true to the mathematics as much as possible.

Still a warning: metaphors are used because we cannot understand the world of the very small in many more dimensions than we are used to. These metaphors give a picture of how to understand these theories. However, they cannot be used to reason in order to find new physics. You will find that some of the metaphors do not exactly match others. What you experience then is the breakdown of the metaphor as a useful tool to understand the theory.

Nevertheless, I hope this book will offer you a world of insight into the building blocks of our universe. It has been my life's goal to understand these things and as far as I have been able to get, I am pleased to share my personal understanding with you and hope you will find some satisfaction in these pictures. Of course, I have stretched these views as far as possible in order to offer the best insight, so if any of these insights turn out not to be (entirely) right, the fault is all mine.

I especially hope to serve those amongst you who find it troubling to understand the complicated mathematics involved, as many people do. I think you have the right to benefit from these scientific results, without having to become a magician yourself.

In the end I wrote the book that I would have liked to read myself.

Good luck & have fun!

Chapter 2

Particles or Waves?



Quantum mechanics shows the peculiar nature of the world around us. It relates to an ancient discussion: is nature built from particles or waves? Or both? Newton preferred a particle view. He discovered that light consists of different colours, and that these colours can be recombined into white light. He explained this behaviour using a particle view of light, which he described in his publication “Opticks” (1704). He managed to explain reflection and refraction using a particle model. He also discovered the so-called Newton rings, which are an interference phenomenon.

At the time of Newton, Christiaan Huygens was more a wave kind of guy. In his “Traité de la lumiere” (1690), he explained light using a wave model based on the principle of Huygens-Fresnel. He created a theory of the polarization of light based on waves. He also explained sound as a wave. Using his wave model, Huygens could explain interference phenomena more easily than Newton.

Electromagnetism was later described by Maxwell and others as a field in which the wave nature of light shows up. In the nineteenth century, the wave-based theory for light became widely accepted. But then, around 1900, Max Planck postulated that light was absorbed and emitted in lumps, each lump with a characteristic energy linked to the frequency of the light. Einstein used this idea in 1905 to explain the photoelectric effect.

The photoelectric effect describes the process of freeing an electron from a conductor by using light. It takes a specific amount of energy to free the electron, and only light of a minimum frequency could do that, as experiments showed. So apparently you cannot “sum up” the energy of light of a lower frequency (e.g., by letting it shine long enough on the electron) to “gather” the energy needed to free the electron. The frequency of the light determines how much energy is transferred to the electron. A lump of energy is associated with each frequency, and only this lump can be absorbed. It cannot be absorbed in part, only as a whole. Clearly, these lumps sound suspiciously like particles. You can compare this with kicking a ball over a hill. When you do not kick hard enough, the ball will roll back. Only when you give enough energy in one kick will the ball go over the top of the hill.

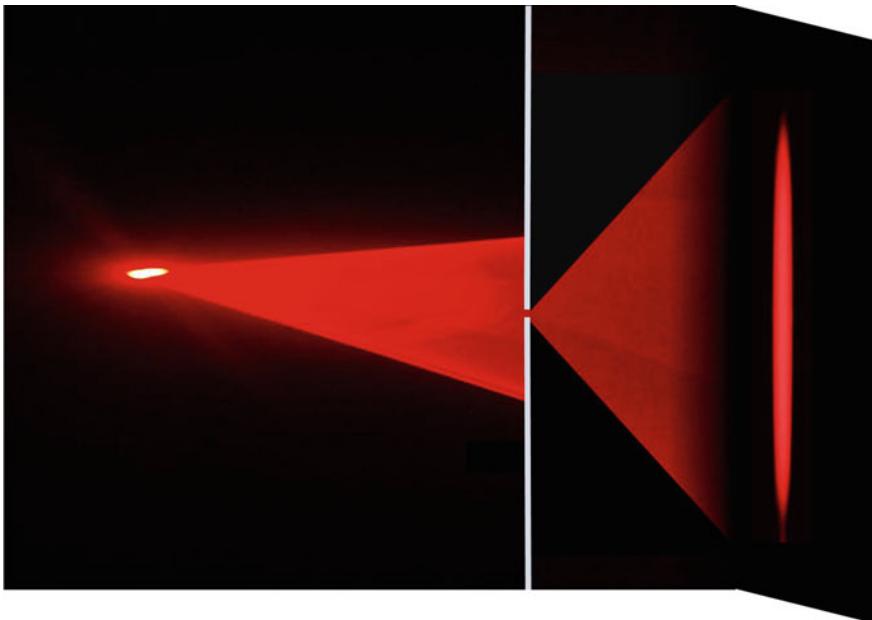


Fig. 2.1 When light passes through one slit in a thin plate, diffraction around the corners of the slit makes the light spread over an angle. The result is a wide blob on the screen behind

The photoelectric effect can be seen as re-opening the debate on the particle nature versus wave nature of light and matter.

The wave and particle characteristics of light come together in the double slit experiment. In this experiment a light source produces light of one colour in all directions. The light hits a wall that has two slits in it. Behind this wall a screen absorbs the light and changes colour where the light hits the screen (like a photographic plate). First, we close one slit, so the light can only pass through one slit. On the screen behind it we see a blob of light (Fig. 2.1). Now we close the slit and open the other slit. Again, we find a blob of light on the wall behind it, but at a different spot. All this can be expected.

What would happen if we open both slits? We could expect just two blobs. But when we open both slits and look at the result, we see something else: an “interference pattern” (see Fig. 2.2). This consists of a bright blob in the centre with many blobs next to it on either side.

This result is a consequence of wave behaviour, as we will explain later. So is light made of waves? Well at first sight it looks that way. However, when we turn down the intensity of the light to a level that we would describe as darkness, while still emitting light at the one wavelength that we can create at the source, we start to see that the light arrives at the screen in spots: one spot at a time. So, every second or so one spot shows up at the screen. Never half a spot, a smeared spot, or more spots at the same time with a lower energy. Always one spot with one energy.

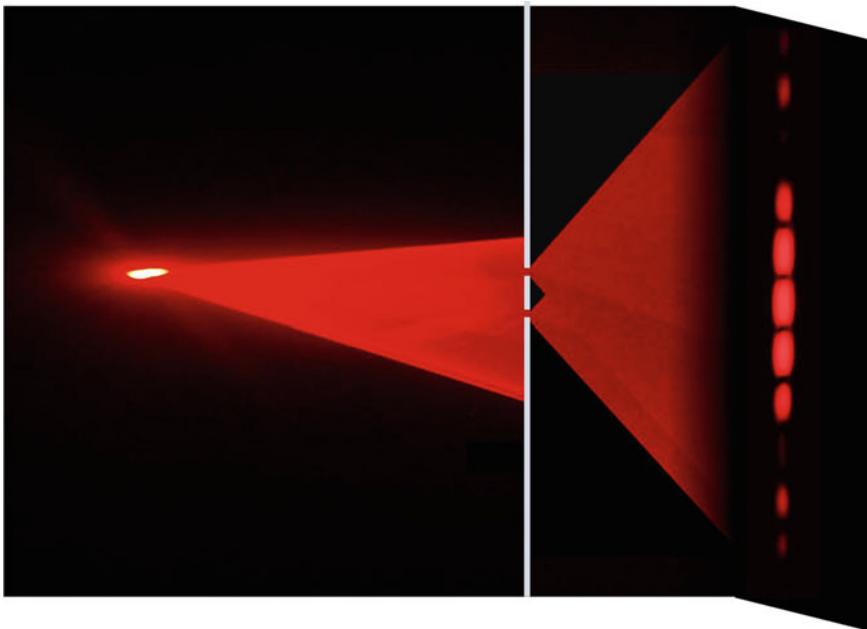


Fig. 2.2 When the light passes through two slits, the light from one slit interferes with the light from the other. This results in an interference pattern on the screen behind. When the light is of just one wavelength, this comes out particularly clearly on the screen

For a given frequency of the light, just one energy is associated, and that energy can be absorbed or not, but never half absorbed, or in part, or smeared out. This is the particle behaviour from the photoelectric effect which leads us to think of light in terms of particles. In this case, a particle of light is called a photon. The fun part of this is that, if we wait long enough, until we have gathered enough spots, we find that they build up an interference pattern. So, in the double slit experiment we see the particle and wave nature of light combined.

Things get stranger when we start to use electrons instead of light. *They show exactly the same behaviour.* So, electrons too are both wave and particle. When we issue one electron at a time and we wait until a lot of electrons have been gathered at the wall, we find an interference pattern, just as with the spots of light. The electrons are produced one by one, they go through the slits as waves, and they get absorbed as particles. How strange that is. The electron cannot interfere with the other electrons that go through the other slit, since each electron is all alone when it goes through! So, the electron must split up somehow, go through both slits at the same time, get together again, interfere with itself as a wave, and finally get absorbed as a particle. How can that be?

Over the past century this has been a puzzle to physicists. Today we can provide some reasonable answers to this question. It involves understanding the world from a wave perspective and it involves concepts such as quantum decoherence. Over the

course of this book, we will gradually unveil the concepts that when brought together provide the answers to this problem.

2.1 How to Describe a Wave

Before we can start to understand how to picture our world in terms of waves, we need to understand what a wave is. How does a wave contain and transport energy? How does it contain and transport momentum? Only when we know this can we start to investigate the answers to our tantalizing problem of wave-particle duality. As we talk about waves, we will explore the question of what is actually waving? This will bring us to the concept of fields and how a field can carry a wave. Then we will finally be ready to understand how a particle can be a wave.

2.1.1 *Wavelength Represents Momentum*

When I punch a spring, it will contract and this contraction will move through the spring until the end of it, where it will hit anything positioned at its end with the same force that I put in (assuming no energy has been dissipated on the way). So, the punch I put in is transported through the spring to its end. Here we see that momentum (the punch) is transported as a wave by the spring.

When I punch the spring, it will start to transport the punch from the first moment I hit it. And as long as I push it in (as a consequence of my punch strength), the pushing will keep folding the spring. Meanwhile, the spring keeps transporting away the fold as a wave through the spring. Now suppose I hit it three times as hard. It will take one-third of the time to push it in, so it will take one-third of the time to fold. This means that from the point of first touch to the moment of maximum fold takes one-third of the time. In that time, the fold could only be transported one-third of the way compared to the previous punch. This in turn means that the width of the fold is one-third as big. It is just more concentrated! So, we see that three times the momentum (punching three times as hard) results in one-third of the wavelength (width of the fold). See Fig. 2.3.

Mathematically, we say that the momentum P carried by a wave is inversely proportional to the wavelength λ :

$$P \sim 1/\lambda$$

The “~” sign means “proportional to”. So, we see that momentum is related to the energy stored in a spatial distance. The shorter the distance the energy of the punch is stored in, the higher the momentum. This means that we could relate space (wavelength) to momentum.

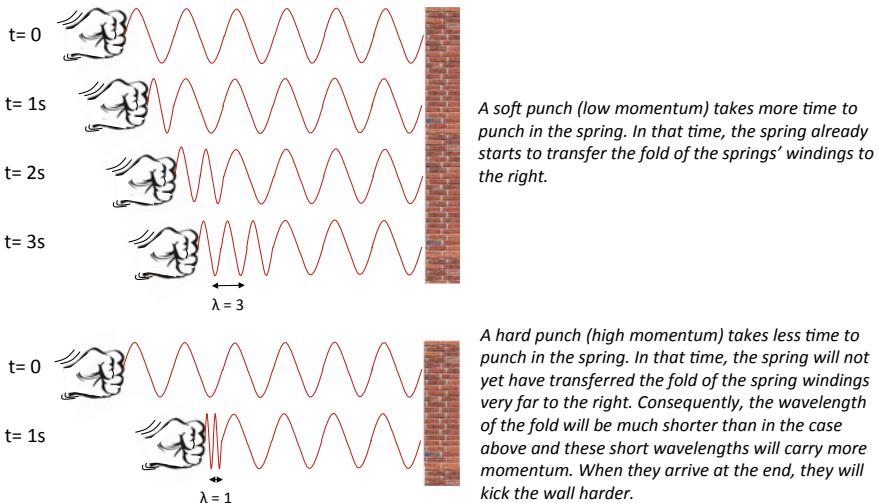


Fig. 2.3 The effect of a soft or hard punch on a spring. A hard punch creates a fold in the windings of the spring that has shorter wavelength, but stronger impact. So, the higher the momentum in the wave, the shorter the wavelength

The type of wave we have been looking at is typically a pressure wave. This type of wave consists of a pressure being propagated through the medium. We call that a longitudinal wave. For such a wave the wave amplitude is in the direction of its propagation. In this example, a punch being propagated through the spring. We can also think about air pressure in the atmosphere. Pressure differences give rise to wind, which is really pressure being propagated through the air. We all know that wind can exert a momentum and carries energy.

However, does this apply to transverse waves too? Transverse waves have an amplitude perpendicular to their direction of propagation. Take, for example, waves in water. Consider a surfer at the beach. When he is surfing a wave he gets momentum from the wave by riding on its slope. The steeper the slope of the wave, the more momentum. When the wavelength of the wave gets shorter, the slope of the wave gets steeper. So here too we see that a shorter wavelength corresponds to a higher momentum, which can be carried over a distance and then transferred, e.g., to a surfer.

You might argue that the surfer gets its momentum (and energy) from gravity, as it is gravity that pulls the surfer down the slope of the wave. However, before gravity can pull the surfer down, the wave needed to push the surfer up. How fast this is done is determined by the steepness of the wave's slope. So, the wave delivers energy to the surfer by pushing him up and the momentum the surfer gets depends on the steepness of the wave.

For those who are not convinced: you find the highest waves in the middle of the ocean. They can easily be 30 m high. But they are not as steep, so a surfer cannot benefit from such waves.

2.1.2 Frequency Represents Energy

Let's take a look at how a wave changes over time. We do this in a simple way by considering one spot in space and seeing how a wave that moves along that spot changes the amplitude there. Imagine you lay a rope straight on a table and pull the left end of the rope up and down in quick succession. The bump you create this way moves at a certain velocity to the right end. Now pick an arbitrary point somewhere in the middle of the rope and paint it red. Then observe what happens to the red dot on the rope when the bump passes by. At first, it lays still on the table. Then when the bump approaches from the left, it starts to move up. As time goes by, the red dot moves further up, continuing to do so until the moment the top of the bump passes the red dot. At that moment the red dot reaches its highest amplitude above the table. Then, as the bump moves on to the right, the red dot starts to go back down until, after a short time, the bump has gone right past the red dot and it lies back on the table.

Suppose we make a graph with time on the horizontal axis, and the amplitude of the red dot on the vertical axis. What does it look like? It looks like a bump itself (see Fig. 2.4)! So, when we consider a point in space and a wave passes by, the point in space actually sees its amplitude go up and down and up and down. The point experiences a wave in time. How fast it will go up and down is determined by the velocity at which the wave goes past, but also by the wavelength of the wave. If the

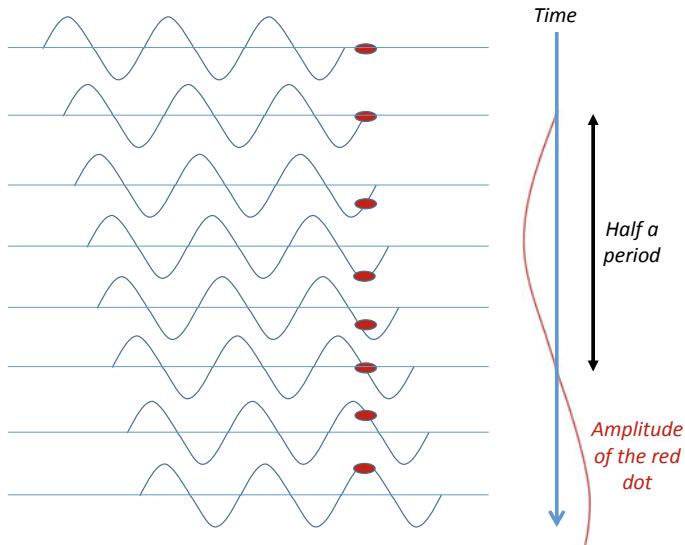


Fig. 2.4 The red dot on the rope goes up and down when a wave passes in time. When the wave passes at higher velocity, it will go up and down faster: it has a higher frequency. When the wavelength of the wave is shorter, the dot will also go up and down faster and its frequency will again be higher

wave goes by faster, the red dot will go up and down faster (its frequency will be higher). So, the frequency is proportional to the velocity of the wave. When the wave has a shorter wavelength, the red dot also goes up and down faster. Therefore, the frequency is inversely proportional to the wavelength. In short, the frequency of the red dot increases when the wavelength gets shorter and/or the velocity of the wave increases. We can summarize this in the following important formula for waves:

$$\text{Frequency } F = \text{velocity } V / \text{wavelength } \lambda$$

Frequency is related to energy. If we think about the rope again and we want the frequency of the red dot to be higher, all we have to do is pull the rope up and down faster. This requires more energy. If we really want a very high frequency, we need to pull the rope up and down like crazy. If we do this, we find that the faster we want it to go up and down the more energy we have to put in. So, we can say that the frequency is proportional to the energy we put into the wave.

At the same time, we saw earlier that the wavelength is a measure of the momentum of the wave. When the momentum goes up, the wavelength gets shorter. Clearly, when the momentum goes up, the energy also goes up. All this can be put into the following equation:

$$\begin{aligned} \text{Energy } E &\sim \text{Frequency } F = \text{velocity} / \text{wavelength } \lambda \\ &\sim \text{velocity} \times \text{momentum} \end{aligned}$$

For particles in classical mechanics we have $\text{energy} = \frac{1}{2} \times \text{mass} \times \text{velocity}^2 = \frac{1}{2} \times \text{momentum} \times \text{velocity}$, since $\text{momentum} = \text{mass} \times \text{velocity}$. So, when we take the frequency of a wave to be its energy and the wavelength to be a measure of its momentum, we get the same type of relation between momentum and energy for waves as for classical moving particles. The formulas differ by a factor of $\frac{1}{2}$. This is a consequence of how waves have to be grouped to look like a particle. We will get to this later.

We can conclude that a wave can be attributed the characteristics of energy and momentum, just like a particle. Let's see how this works for photons, the "particles" of light. A *wave* of light would obey our formula:

$$\text{Frequency of the light } F = \text{velocity} / \text{wavelength} = C / \lambda$$

where C is the velocity of light. Now let's insert the way energy E is related to frequency ($E \sim F$) and momentum P is related to wavelength ($P \sim 1 / \lambda$). Then we get:

$$E \sim F = C / \lambda \sim CP$$

The formula $E = CP$ is the relation between energy and momentum for (massless) photons in the theory of relativity. And so, we see that our understanding of

momentum as related to the wavelength and energy as related to frequency gives waves just those properties we are used to from the particle world. This again suggests that particles could be represented by waves. Of course, this does not yet give us any understanding of how photons can be quanta. So far, we only understand how frequency is related to energy and wavelength to momentum, and that this gives us the right equations.

Let's go back to the beach. You might be aware of the fact that wavelengths in water get shorter once waves start to close in on the beach. This would mean that their momentum goes up, and this is actually true! But you might wonder, how can that be? Isn't momentum something that is conserved? Yes, it is! The point here is that the total momentum in the sea below the surface does not change. But when the wave gets into shallower water, the momentum is spread over less water depth. So, the wave at the surface starts to carry more of the total momentum.

Another issue is that the total energy must remain the same. This means that the frequency must remain the same despite the wavelength getting shorter. The only way to do this is by lowering the velocity (remember, $F = V/\lambda$). And that is exactly what happens. The velocity of the waves in the water declines by the same amount the wavelength gets shortened. We will use this idea later when we discuss what happens to waves that enter a different medium.

2.1.3 Superposition and Interference of Waves

Waves have properties that we do not see in particles. Superposition is one of those properties. Superposition means that when two waves pass a certain spot, the amplitude at this spot is the sum of the amplitudes of the individual waves there. Note that the amplitude is negative for the part of a wave below the average level.

Two waves can amplify each other when they are “in phase”, meaning that both waves have the same amplitude at a particular spot (see Fig. 2.5). Adding up those amplitudes creates a higher resulting amplitude. On the other hand, two waves can cancel each other out when they are in opposite phase, meaning that the waves have opposite amplitude at a certain spot. Adding a positive amplitude to a negative

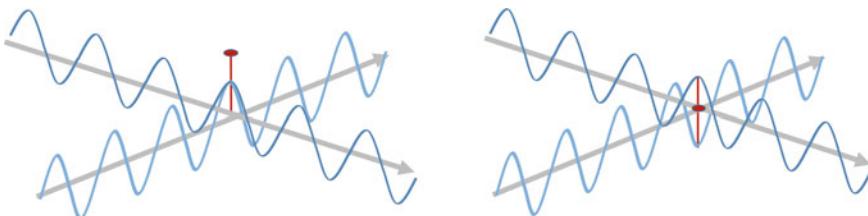


Fig. 2.5 Left: two waves in phase at the red dot will add up to a higher amplitude. Right: two waves in opposite phase at the red dot will cancel out. The resulting amplitude is zero in this example

one creates a lower resulting amplitude. This is what happens in a noise reduction headphone. It measures the waves approaching the phone and tries to create an opposite wave that would cancel out the incoming wave in the region of your ears.

When waves meet each other, the superposition of those waves creates a pattern. This pattern is called an interference pattern. This pattern shows up everywhere, where the waves interfere with each other. Take for example a water wave. When the wave hits a wall with two openings in it, each opening acts as if it were the source of a new wave. This leads to two waves coming from the two openings (see Fig. 2.6). A little further along we see that the waves start to interfere. Each blue line in the wave represents the top of a wave. Where the lines cross, the waves interfere positively to an amplitude of twice the amplitude of each individual wave. Where a line of wave number one crosses the midpoint between the lines of wave number two, they cancel each other out.

The interference between the two waves comes out most clearly at the right wall. Here we see that some parts of the wall experience a very high amplitude, while other parts experience hardly any amplitude.

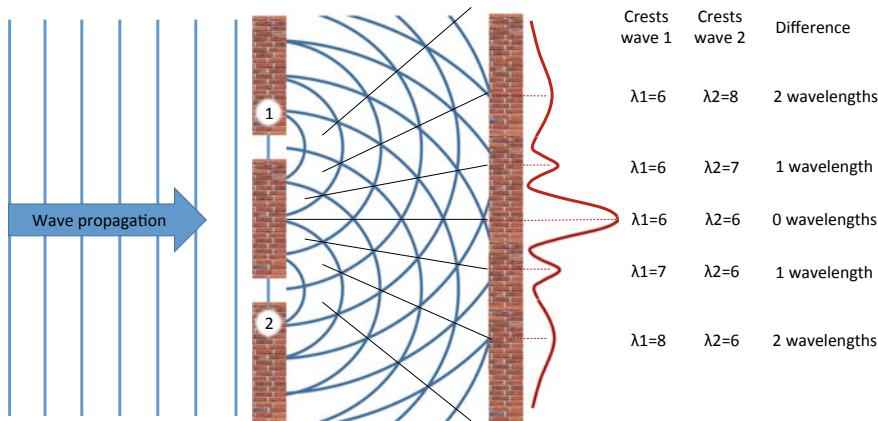


Fig. 2.6 Interference pattern of two waves coming from opening 1 and opening 2 in the left wall in the water. The blue lines are the crests of the waves. The straight black lines show where the two waves interfere positively. The right wall experiences the interference pattern between the two waves as shown in the red graph next to it. So, where the red graph shows a high amplitude, the water is making high waves against the wall. If one counts the number of wavelengths (= the number of blue lines) from each opening on the left towards the right wall, one gets the numbers on the right of the picture. When these numbers differ by an integer number of wavelengths, the waves interfere positively and we see a maximum in the red graph. In between, the waves interfere negatively and we see a minimum in the red graph. At these points the waves differ by an integer number of wavelengths $+ \frac{1}{2}$

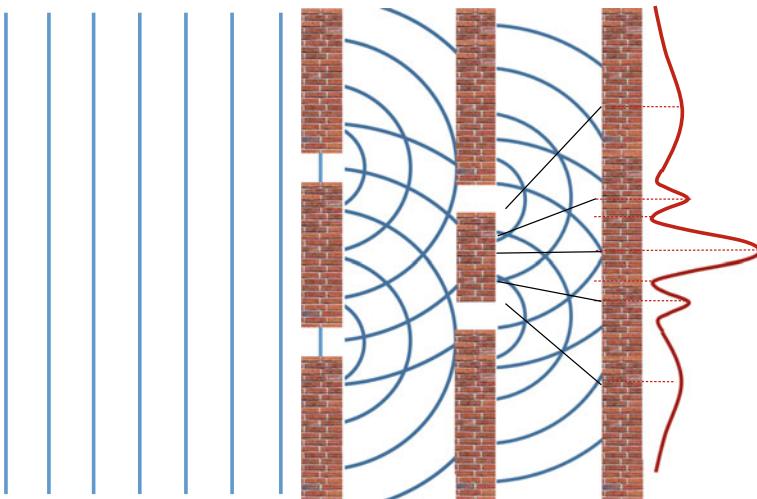


Fig. 2.7 When a new wall is put between the other two, with holes at a different spot, the interference pattern is determined by the waves interfering between the middle wall and the wall on the right. So, waves interfere only in the region where they can travel undisturbed

2.1.4 Measurement

The interference between the two waves in Fig. 2.6 takes place as long as they are undisturbed. What would happen if they were disturbed? For instance, if we put another wall with holes in the wave path, the interference pattern at the wall behind would change and would be determined by the path of the waves between the new wall and the wall on the right (Fig. 2.7). So, the interference pattern is determined by the path of the two waves between two walls, i.e., between two events. Each event (each wall) changes the interference pattern.

In the world of particles, we will see that such an event is equivalent to a measurement. A measurement shows us for instance where the waves are located (at the openings in the wall). At the same time, unavoidably, the measurement (the wall) changes the shape of the wave. We will see later that this is a very important property of all measurements. A measurement gives us the opportunity to get information about a wave, e.g., about its position in space–time. But, while doing so, it has to change the wave. During the time when no measurement is being made, the wave has a chance to develop. During that time waves can interfere with each other. So, if the wave on the left of Fig. 2.7 represents a particle wave, this wave gets split into two waves at the wall's openings. The interference between a particle's waves takes place for as long as they are undisturbed, meaning for as long as they are not measured.

A measurement must involve a *transfer* of momentum and/or energy that limits the state (e.g., position) of the particle. Only then can that state be measured. Let's

look a little further into this statement. For instance, would it be possible to measure the position of a particle without changing its momentum? In Fig. 2.8 we find two particles on the left that interact with each other, but this has not led to a change in momentum. Both particles just fly straight through. If I were to measure particle A I would have no way to deduce where it interacted with particle B. It could be anywhere along the route it has taken. That means that I have no idea about the position of B at the time of the interaction. On the right we find two particles that have interacted and repelled each other. During the interaction the momentum changed in direction (and possibly in value). Measuring particle A will now give me an idea of the centre of interaction. Hence, I can derive the approximate position of particle B. So, this picture clearly shows that a change in momentum is required to measure a position.

Let's look at another example, in which a light bulb indicates where an electron is in a double slit experiment. The light would transfer momentum to the electron. This would happen at one of the holes. The result is that the light path would get bent by the electron at one of the holes. So now we can distinguish which hole the electron has gone through! This limits the position of the electron to that one hole. The two paths (the two holes) are no longer indistinguishable. But the interaction with the light also changes the momentum of the electron. When we do this, the interference pattern disappears. What we would like to do is to trick nature into showing which hole the electron went through *and* yet still see the interference pattern. Suppose we lower the momentum transfer, so that the interference pattern would show up again. How can we do that? We can make the light's wavelength longer, since that would mean the momentum of the light was reduced. Consequently, the light would transfer less momentum to the electron. Indeed! The interference pattern returns. But now we have another problem: the wavelength of the light has become longer than the distance between the slits. That is a problem, because the light flash can only show us where the light came from with an accuracy of about one wavelength. So now we get

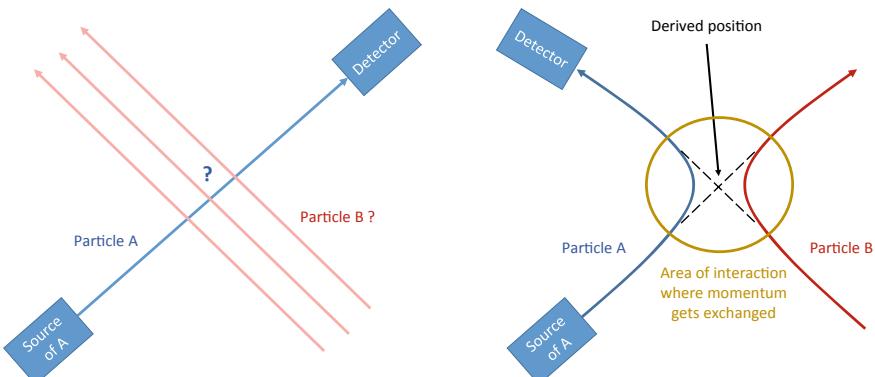


Fig. 2.8 Only when two particles exchange momentum (on the right) can we derive information about the position of the interaction point and the approximate position of particle B at that time. When momentum does not get exchanged (on the left), we cannot get information about the position of particle B

a flash when the electron passes through, but we can no longer make out which slit it went through. So apparently, the wave properties (interference) only exist within the area of space where we cannot make out the position of the wave/particle. This is the same as saying that the wave properties only exist for all “quantum states” of a particle that are *indistinguishable*.

This behaviour has serious consequences in particle physics. If we want to probe something in more detail (i.e., in a smaller region of space), we need to put in more momentum. In Fig. 2.8, the particle that does the probing, particle A, needs to have more momentum in order to make the interaction area smaller.

So, the part of the sentence “...that limits the position of the particle” is relevant. Every transfer of momentum and/or energy will change the interference pattern. But a small transfer of momentum will not change it enough to make the pattern disappear. Beyond a certain amount of momentum transfer the pattern will change significantly and we can start to speak of a measurement event. So, in the case of the example above, this means a strong enough light of a short enough wavelength. Absorption by the wall is also such a ‘high enough momentum transfer’. The wall that we inserted absorbs some of the momentum of the waves and limits the position of the particle, so it is therefore relevant in changing the interference pattern. Summarizing, a measurement of a state implies that you exchange enough momentum and/or energy to be able to distinguish between states, e.g., if you want to distinguish which hole the electron went through (distinguish between two position states). As soon as you can distinguish between states a particle can be in, these states will no longer interfere as a wave.

2.2 Probability Amplitude

Let’s go back to the question of how the electron splits up, goes through both slits at the same time, gets back together again, interferes with itself as a wave, and finally gets absorbed as a particle. How can that be?

This paradox has puzzled physicists for a long time and still does. A prevailing opinion is that the wave character is just a probability. How does that work? It starts with a very important principle in quantum mechanics: if you want to calculate the probability of something happening, you need to include all the possible ways it can happen. That makes sense. Suppose you want to calculate the probability of throwing a total of 7 points with two dice, you need to include all the possibilities that give “7”. So, $3 + 4$, but also $4 + 3$, $5 + 2$ and $2 + 5$, $6 + 1$ and $1 + 6$. The sum of the probabilities for each of these combinations gives the total probability of throwing 7.

But that is not all. If we look at the double slit experiment, just adding the probability of going through slit A to the probability of going through slit B does not give an interference pattern. It just gives two blobs on the screen behind. So, there is a special recipe for adding these probabilities in quantum mechanics. That recipe requires a wave.

Let's take for example a particle starting at the left of Fig. 2.6. Let's call that point A. We want to calculate the probability of ending up at one particular point on the wall on the right. Let's call that point B. First, we have to assign a wave to each path from A to B. What the wave looks like is entirely determined by the characteristics of our particle. When both slits are open, there are two indistinguishable paths. Now we have to determine how the waves add up (superposition). So now we have some resulting amplitude for the interference pattern at point B. That can also be a negative value, e.g., when the wave is in a trough. However, a negative value cannot represent a probability. Hence, we need to take the absolute value squared of the amplitude to ensure that we always get a positive value. This value is the probability of arriving at B.

The paths in this example are indistinguishable since we do not have any means to detect which slit the particle went through. If we were to install such a means (the light bulb), the two paths would no longer be indistinguishable and the calculation would be based on only one path through one slit. The result would be a blob behind the slit the particle went through and no interference pattern.

It is clear that this recipe gives us the right probability distribution. The recipe is called the Born rule [Ref. 1] and is used for all situations that cannot be distinguished in quantum mechanics. For instance, if we have two identical particles at close range, we can only describe them when we include the probability that they switch places. This situation is completely indistinguishable from the unswitched situation. We will see later what the consequences are when we have two indistinguishable fermions or two indistinguishable bosons. Similarly, when we want to calculate the probability for a particular interaction to take place, we need to determine the superposition of all possible indistinguishable variations by which that interaction can take place. The probability for that interaction to take place is determined by the superposition of all these possibilities.

The fact that we have to include all probabilities of getting from A to B is easy enough to understand (as with the dice). But why does this sum have to be calculated using a wave? We know that the recipe works since it agrees with measurements, but that does not tell us why. The resulting amplitude of the superposition of interfering waves is called the “probability amplitude”. This probability amplitude tells us what paths will be travelled most frequently and what paths are very unlikely to be travelled.

So how should we understand such a probability amplitude? In the Copenhagen interpretation of quantum mechanics, the probability amplitude had no physical meaning. It was just a recipe to calculate the probability of finding a particle in a certain state (e.g., at a certain point in space–time). That approach carefully avoids having to think about the problem. But I would say that the probability amplitude is too important to just avoid thinking about its meaning. So, let's give it a try.

Suppose we consider the probability amplitude as a means to carve out paths for the particle to move along. This way we would combine the wave character of the probability amplitude with the particle character of the particle itself. Following this idea, a single particle that starts at A will just go its way. Supposedly, it does not know anything about the paths, slits, or probabilities. It will move to the right through one of the slits (we don't know which!) and end up at some point on the wall on the

right. When many particles do this, we see that most particles end up at a spot where we have calculated a high probability amplitude. However, each individual particle only takes one of the possible paths. So, when there are many paths taking particles to a high amplitude spot on the right and there are only a few paths towards a low amplitude spot, we can understand that the ignorant particles will end up mostly at the highly probable spots. We could view the probability amplitude as the roads in a country. Many roads lead to the nearest city, and only a few towards the less inhabited lands outside the city. Consequently, it is no surprise that most cars end up in the city. The probability amplitude describes the roads. The particle is like a car.

But for me, this is not very satisfying. We need to assign a wave to a particle and use that to determine what the probable paths are. The wave is determined by the characteristics of the particle, amongst which there is its momentum. So how does the path get carved out for the particle? If the slits are moved to a different (relative) position, the interference pattern changes. So, the wave that carves out the path must have some dynamical knowledge about the particle as well as the positions of the slits and many other things. What could such a wave be?

So, the problem comes down to the question of *why and how would the paths be carved out according to such a probability distribution, resulting from a superposition of waves?*

2.3 What Is Waving?

In order to understand our question better, let's first look at the problem of causality. Causality says that there is a link between cause and effect. In the present case, the cause is the properties of the particle and the environment and the effect is the carving out of the path for the particle. One thing we must understand about causality is that *causality is not emergent*: cause and effect cannot emerge out of nowhere without being physically linked in some way or another. At least, no scientist has created a theoretical universe yet in which causality emerged without putting it in first. Hawking, amongst others, hoped that causality would emerge from quantum fluctuations that do not carry such causality intrinsically. However, computer simulations in a theory called “causal dynamical triangulations” can produce a viable universe only after causality has been put in explicitly [Ref. 49, 50]. Consequently, we must be able to distinguish between cause and effect. This means that there must be something physical linking the effect (possible paths of a particle) to its cause (characteristics of the particle and its environment). Only then can we be sure that a change in momentum or a change in the positions of the slits (the cause) will result in the changed interference pattern (the effect).

One option is that there is something physical continuously carving out the path for the particle. If that were so, that something must be connected to both the context of the particle and the particle itself. It must connect dynamically: at each moment in time, it must reconsider the path depending on what changes in the environment as well as in the particle itself. This seems a very complicated way to view the situation.

A simpler view would be that the particle does not follow some magical probability distribution that gets carved out for it, but actually determines its own way. But then, how can a particle have knowledge about its environment? Again, *the simplest solution would be that the particle actually is a wave*. We know that waves show exactly the right characteristics. An electron that is a wave would go through both slits, interfere with itself and produce an interference pattern. Using this picture, we do not have a separate particle that has to “communicate” with its probability distribution. There is no problem with cause and effect between the particle, its environment, and the probability distribution. When a particle is just a wave, the wave goes through its environment, interferes, and comes out with the right probability distribution.

But then, how can energy be absorbed in lumps? After all, the waves get spread over the wall behind, so what decides that the electron gets absorbed in its entirety in one place instead of the other? Don’t you just hate that? Never a break!

Currently, many researchers are considering alternatives. An interesting one is that the wave is regarded as a “pilot wave” that carries the particle [e.g., Ref. 64]. However, this idea once again requires two entities somehow operating together. So before introducing extra stuff, let’s see if we can produce particle-like behaviour *just from waves*. This is the main idea behind quantum field theory. This theory is the most successful theory in describing particles, light, and forces and gives results that agree with measurements to very high accuracy.

So, in order to understand the world, we will need to dive into the behaviour of waves and see how these can form lumps of energy. We will also need to understand something about the “stuff” that does the waving. In quantum field theory, this mysterious stuff is referred to as a field. Quantum field theory investigates and describes the (quantum) characteristics of fields. This theory is a highly mathematical theory. In this book we will explain how fields can be understood step by step in a conceptual way. The concepts link closely to the mathematics, but you will not need a mathematical background to be able to understand the concepts.

Chapter 3

Fields and Waves Making Up Reality



When we think of a wave, we think for instance of waves in water. Or we consider a rope on the table that is excited by swiftly pulling one end up and down, resulting in a wave that moves towards the other end of the rope. So, to get a wave we need something that we call a medium: water or the rope. The medium carries the wave.

We saw earlier that something must be waving in order to get interference. Now we see that we need a medium to carry such a wave. What could that be? Quantum field theory says that a wave is a disturbance or excitation of a field. But that field must still be carried by a medium. This does not make clear what that medium actually is. We simply do not know, but we will call it the vacuum. We just know that it must be there, and that it can carry a wave.

3.1 What Is a Field?

By appeal to analogy, let's take a look at the atmosphere as a medium. The atmosphere can carry all sorts of waves. An obvious one is sound waves. When I make a noise, someone else can hear it. In the meantime, the sound wave moved from me to the other person. However, on the moon I cannot do this. When I make a sound on the moon, no-one else will hear it. So sound waves are bound to the medium. When there is no medium, there cannot be any waves. In the atmosphere, I would be able to measure what the sound level is at any point (i.e., what sound waves are going through that point). In theory, I could do this throughout the atmosphere and create an overview of all positions in the atmosphere and what the sound level was there. Maybe I could make a nice graphical picture that shows the sound level everywhere. As the sound level can take a value anywhere in the atmosphere, it is called the sound field. With any type of wave in the atmosphere, I can associate such a field. More generally, a field is anything that can assume a value in every part of the medium.

For example, the atmosphere has a temperature at every moment and at every spot within the space that is occupied by the atmosphere. This is called a scalar field:

the atmosphere can assume one temperature value at every point in space–time. However, the atmosphere can also assume a velocity at every point in space–time. Obviously, this is what we call wind. With every point in the atmosphere, we can associate a velocity and direction of the wind blowing there. This is called a vector field. A vector is often depicted as an arrow with a length and a direction. It differs from a scalar in that it has a direction, while a scalar is just a number without any direction.

A vector field in the atmosphere is three dimensional. It requires three dimensions to describe all possible directions the vector might have. Wind can blow east–west (one dimension), north–south (second dimension) and up–down (third dimension). The latter occurs for example in a thunder cloud.

Here we see one medium (the atmosphere) that supports two fields: the temperature field (scalar) and the wind field (vector). We can find other fields in the atmosphere, e.g., a moisture field (scalar) and a pressure field (scalar).

Another example of a field carrier is a water surface. The water surface literally carries waves. The field value in that case is the relative height of the water above the average. These are typically transverse waves. Water waves constitute a two-dimensional field since they only appear on the surface. However, in the sea we can find pressure waves and underwater streams. So, the whole sea is a three-dimensional carrier of, e.g., a water flow field. This is a vector field, since it has a strength (value) and direction. Pressure waves are typically longitudinal waves.

A solid can also be a medium for fields. One can have waves in solids, and also temperature fields, stress fields, etc. When the stress in a solid is different from point to point, one needs to describe the direction of the stress at each point in the solid as well as the direction in which it changes. Consequently, one needs a matrix rather than a vector at each point in the solid. Such a matrix is called a tensor and the field in the solid is a tensor field. The word “tensor” originates from Latin and means “that which stretches”, or in short tension or stress.

A specific example of a solid is a thick rubber stick. You can imagine twisting the rubber at one end. That twist creates a tension that is propagated through the rubber stick. In fact, it propagates like a wave. The velocity of propagation of such a twist depends on the elasticity of the rubber.

All the types of fields we have discussed here exist in the world around us. They can serve as analogies when we are looking at the vacuum. We will see how different types of particles are actually waves in different types of fields in the “vacuum” medium. For each species of particle, we will have a different type of field. To get a picture of how that works we may use the examples of fields given in this section. However, keep in mind that they are just analogies! Some examples of fields we will meet in this book are:

- The Higgs field is homogeneous (as we will see in the section on symmetry breaking) and has the same strength everywhere. It is a scalar field.
- Most particles of the boson type are excitations in a vector field. There are as many different fields as there are types of bosons. For example. Photons are bosons and

they are excitations (waves) of the electromagnetic field. The electromagnetic field exists throughout the vacuum.

- Particles of the fermion type are excitations in a spinor field. Spinor fields are not easy to understand, so we will discuss them at length in Sect. 13.2.

3.2 All We Are Is Waves in a Field

So, particles are supposed to be waves in a field. Let's investigate what that means.

First of all, this goes for all the “particles” that we are made of: we are all just made of waves in a field. The vacuum is not something empty (a big nothingness), but is a medium that carries fields that are waving all the time. What looks to us like a billiard ball on the billiard table is actually a bunch of waves moving from the queue to the pocket. This is an important picture to keep in mind, since we will build up the way the world works from this picture.

3.2.1 Objection 1: How Can a Billiard Ball Be a Wave?

At this point the idea that we are made of waves may seem strange and hard to accept. For instance, how can a wave be standing still in the way a billiard ball can? You have to imagine that at the (sub) atomic level nothing is really standing still. Electrons can be seen as waves moving around the nucleus of an atom. The nucleus consists of protons and neutrons. These in turn consist of quarks. Quarks are waving around each other. As long as such particles are waving in circles, the circle as a whole can stay at one particular spot in space–time.

We can also assign a so-called de Broglie wavelength to a large object such as a billiard ball. However, the wavelength of a billiard ball is extremely short due to its large mass. It can be calculated using the following formula:

$$\lambda = h/mv$$

h = Planck's constant = 6.626×10^{-34} Js;

m = mass; v = (nonrelativistic) velocity

This formula should not come as a surprise, as we know that momentum $p = mv$, so this formula simply states the relation between momentum and wavelength ($\lambda \sim 1/p$) as we saw before. For a billiard ball the momentum is very high compared to Planck's constant because of its mass. So, unless the mass of the particle is extremely small, the wavelength is always going to be very short. The wavelength of a billiard ball is smaller than the size of an atom. Keep in mind that the billiard ball will never be standing completely still. For instance, it has a temperature and consequently, its atoms are constantly “shaking”. Hence, the atoms of any object will always be

moving to some extent. Therefore, any “normal size” object will never have a zero momentum nor a long wavelength.

3.2.2 *Objection 2: But Things Do Not Look Like Waves*

You might argue that an object such as a billiard ball does not look like a wave. However, what we see from an object are photons coming from the “surface” of the object. The observed photons are translated in our brains into a useful picture of the world. It is useful to know whether an object will be rough or smooth to handle. And so, our brain interprets our observations as “rough” or “smooth”.

In fact, no object is ever “smooth” as it consists of nuclei and electrons that are very small compared to the size of an atom. So, any object is mostly empty space with here and there a nucleus or an electron. If we could blow up a nucleus to the size of the sun, then the electrons would be planets and the atom would have the size of the solar system. What we know of the solar system is that it is primarily a whole lot of empty space.

Another aspect we need to stress here is that things of the size we can see contain a lot of atoms that continuously interact with each other. They also interact all the time with many photons, cosmic radiation, other molecules (e.g., air), etc. Each such interaction limits the position of the object in the sense that the waves constituting the object do not have much chance to develop before the next interaction happens. This is why we primarily see things at a particular position and we experience classical behaviour instead of wave behaviour. This process is referred to as quantum decoherence. We will go into the details of that process in chap. 12, when we have built up a sufficient understanding of the interaction processes involved.

3.2.3 *Objection 3: How Can Waves Make a Table Seem Massive?*

An object may appear as massive to us. It could also feel like that, e.g., a table would be massive. We cannot put our hands through a table. What does that massiveness mean when the table is actually a lot of empty space with here and there a nucleus or an electron? The same goes for our hand, also just empty space. Why can't we put our hand through the table? It seems likely that none of the particles in our hand need ever touch any of the particles in the table, so why not?

The reason lies in the forces between the nuclei and the electrons. The atoms in the table can only form a table because they attract each other. So, it is electromagnetic forces that make the table a consistent object. The same goes for my hand. When I try to put my hand through the table, the electrons in my hand repel the electrons in the table. And since each object is held together, these electrons are not going to get

out of the way. If I push my hand on a table, the atoms of the table hold on to each other due to electromagnetic forces amongst them. They also resist my hand due to the electrons in the table repelling the electrons in my hand. If I push so hard that the table breaks, this means that the forces amongst the atoms in the table were simply not strong enough.

So, we can conclude that the whole reason why I cannot put my hand through the table is because of electromagnetic forces between atoms and electrons! As we will see later, these forces are also made of waves. Like the electromagnetic force that is made of electromagnetic waves. What we experience as massive is in nature only waves and interactions.

3.2.4 Objection 4: But Waves Die Out, Don't They?

In the media we know such as water, ropes, air, metal, or any other substance, waves cause friction. Friction is essentially loss of energy. You may picture this as the wave creating other waves in the medium that cause the temperature of the medium to rise a bit. So, the energy gets spread out.

In the vacuum we will see that classic friction does not appear, but other processes cause the waves to create other waves. The difference is that these waves either recombine with the original wave, or die out themselves and hence do not carry energy away. The waves we find in a field in the vacuum are quantized (see Chap. 7) and cannot dissipate energy. Consequently, a wave in vacuum can exist and maintain its consistency over a long time, so even after a billion years an electron wave is still an electron wave.

3.2.5 Objection 5: What, Then, Is Empty Space, Through Which I Can Throw a Ball?

Empty space does not exist in the way we think we know it. Space–time is filled with the vacuum, but the vacuum is a kind of stuff that is a medium for the types of waves we call particles and forces. Try to picture it as something like the atmosphere. It fills the space around us on earth. Remember that we could create sound waves in the atmosphere, but we cannot do that on the moon. The atmosphere is an example of a medium we can step out of and we can experience the difference between inside or outside the atmosphere. If we could step out of the vacuum, the waves we call light, particles, billiard balls, or tables could not exist since there is no medium to carry them, just as creating a sound on the moon is not possible.

So, the vacuum is all around us, on the moon (and in the moon, otherwise the moon itself could not exist), in and around the ball we throw, between the stars, and

basically everywhere in the known universe. So, there is no such thing as stepping out of the vacuum. You could view it as a never-ending atmosphere for sound waves.

Since space is filled with the vacuum, it is not empty space. Actually, when you think about it, empty space would be a strange thing. Just go back to the idea of a ball we can throw through empty space. Imagine we do this in a “truly empty” space with no gravity. We throw the ball and it will keep on going in a straight line (Newton’s law). How does the ball know what a straight line is? There is no reference. How will it maintain its momentum? There is no reference.

Of course, we know that when no force is exerted, there is no reason to deviate from a straight line so the ball will keep on going in the same direction. The same goes for its velocity. Nothing changes its velocity. But this is not what I mean. What I mean is, how would the ball know what going straight means?

By comparison, if I look at the wave picture, the ball represents an amplitude in the field. This amplitude pulls the next spot in the field. This actually exists, since it is the next spot in the medium (e.g., the next air molecule). And so, a wave can propagate through the medium. It will do so in a straight line, since there is nothing that disturbs the waves to do otherwise, and with the same wavelength (i.e., momentum). So, the medium provides the reference.

Tell me then, how a ball in empty space would do that without a medium? What is actually the definition of the next spot in empty space? In our minds we always imagine a rod or axis to represent a direction in space, and on the axis or rod we imagine measures of distance. But when space is empty, what would provide this reference?

It is actually rather hard to find the empty space picture a strange one. This is caused by the fact that our daily experience (our mind) explains throwing a ball as something flying through space. So having grown up with that, it is not easy to see that this is actually a weird business.

3.2.6 *Objection 6: Oh-No, Not the Ether Again...*

Moving through a medium... we have seen this before. In the late nineteenth century, the prevailing picture was that we all move through the ether. The ether was an all-pervading medium, too. People tried to measure the velocity of the earth relative to the ether. They failed to find any velocity, no matter what time of the year (and the earth moves in another direction each time of the year since it revolves around the sun). So, this was puzzling to say the least.

Then, Einstein came along with the theory of relativity. He said that you can only measure your velocity relative to something else. There is no absolute velocity, no universal frame of reference that has a zero velocity, relative to which you can define your own velocity. The speed of light is always c . But you cannot define your own velocity relative to it, because it is always c , no matter what your own velocity is relative to any other object. This can only be so with the help of some strange effects such as length contraction and time dilation (we will come back to this).

In fact, Einstein considered the constancy of the speed of light to be a law of nature. And as relativity would have it, no law of nature should be different when you change your velocity. This makes sense, since if the laws of nature were to change depending on your velocity, you would be able to figure out your absolute velocity (or velocity 0). Velocity 0 would be when that law has a particular value or property.

So how does this look in the wave picture? First of all, we can only measure ourselves against each other, against other waves! We cannot measure ourselves against the vacuum. We need another wave to interact with to be able to measure ourselves against. So, the picture of the vacuum as a carrier of waves is essentially different from the picture of an ether as in the nineteenth century. In that picture, the ether was expected to be something we would not be able to detect directly, but something we could detect indirectly by measuring our velocity relative to it. The ether was considered to be a kind of stuff that fills the empty space, so it was taken to be something on the same level as matter. The vacuum as a medium is not on the same level as matter. It exists on a more fundamental level and it is not made of anything familiar to us. In fact, we will not be able to discover what the vacuum is made of. By comparison, it is also not possible to use water waves to discover that the medium “water” is really made of hydrogen and oxygen atoms. Waves in a medium will only be able to learn about some properties of the medium they are waving in, but they will not be able to reveal the composition of the medium.

The nineteenth century concept of ether was falsified using the speed of light in an experiment by Michelson and Morley in 1887 (see Fig. 3.1). In that experiment a beam of light was split and sent in two directions. It was assumed that light has a constant velocity in the ether. So, if we have a velocity with respect to the ether, this would be noticeable through the difference in light speed in two perpendicular directions. One of these directions could be aligned with the velocity of the earth relative to the ether, in which case the other direction would be perpendicular to it.

The result of the experiment was that there was no difference in the speed of light travelling in different directions. The experiment has been repeated many times with different instruments by different researchers, and never has a difference in speed been found. The only possible conclusion is that one has to accept that the speed of light is the same for everyone, no matter what velocity we have with respect to light or to others. And no absolute velocity could be determined with respect to the ether.

In the wave picture, c is the maximum velocity at which waves can propagate through the vacuum. As we will see later, the value of c is related to the “elasticity” of the vacuum. In the wave picture, this is what it means when we say that the constancy of the speed of light is a law of nature. Of course, the “elasticity” of the vacuum, the medium in which we are all waves, must correspond to a law of nature.

When I have a velocity (i.e., I am a wave with a certain velocity in the vacuum), the speed of light is still c from my perspective. How does that work? Should the light not go faster through the medium? No, it does not. This requires effects such as the length contraction of my wave when I go faster, leading to measuring the other (light) waves differently. These effects cancel out the difference in velocity between my wave and the light wave when I measure the speed of the light wave. We will

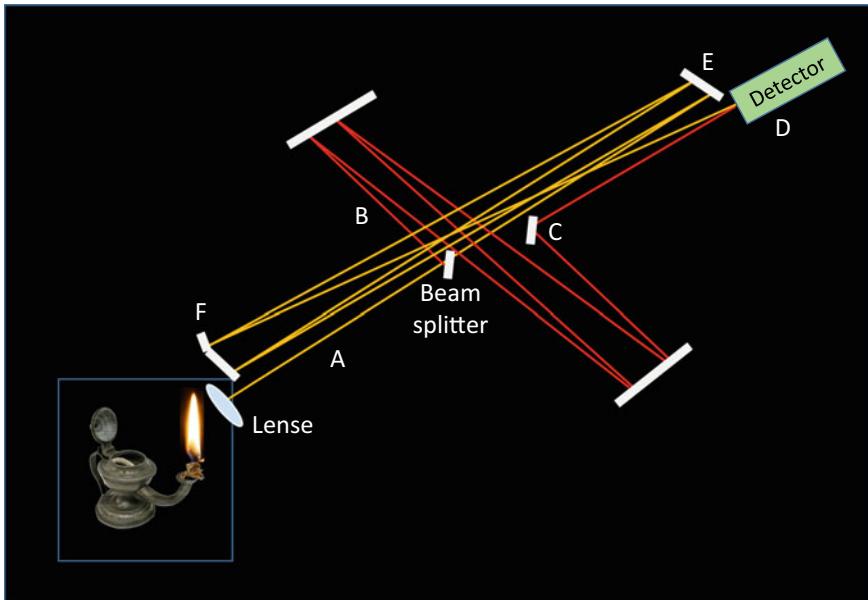


Fig. 3.1 Schematic view of the Michelson-Morley experiment. An oil lamp provided the light that was focussed via a lense into a beam (**a**). The beam was split by a beam splitter into a reflected beam (**b**) and a beam going straight through. The beam (**b**) was reflected a few times until it hit mirror (**c**), which reflected the beam into the detector. The beam going through the splitter was also reflected a few times (mirrors (**e**) and (**f**)) and finally reflected into the detector. So both beams spent most of their time on a path perpendicular to each other. In the detector, the two beams interfere and produce an interference pattern. Slight differences in the speed of the light along the different paths would be detected as a shift in the interference pattern

come back to this later. First, we will investigate waves a little more, and consider what it actually means to be a wave rather than a particle.

3.3 Conclusion

The vacuum is a medium that carries waves. All particles and forces we know of are waves in that medium. We can only do measurements on waves we interact with, and so we can only measure other waves (just as we can hear sounds, but we cannot hear the air itself). We cannot measure the vacuum directly. The vacuum does not provide any other means to define an absolute velocity, and so it is essentially different from the ether as defined in the nineteenth century. The vacuum has a kind of “elasticity” that can be considered a law of nature, defining c as the maximum velocity of waves through the medium. All the waves in the vacuum appear in different fields in the

vacuum. Just as air could contain pressure waves, air temperature, and moisture as different fields carried by the same medium.

Chapter 4

What Is a Particle If It Is a Wave?



So far, we have been talking about waves. We have also seen that typical particle features such as energy and momentum can also be carried by waves. We have claimed that all of us are waves in the vacuum. However, this does not yet tell us *how* a particle can ever be a wave. When we imagine a particle, we see a little ball, located at a particular position at a specific time and propagating at a particular velocity. If particles are to be represented by waves, then we must be able to get these properties from the waves. Only then can we understand how we can see (= measure) a particle while we are actually dealing with waves.

So, in this section we are going to investigate how a particle is based on waves and how that relates to the momentum and energy of the particle. This will lead to a rather interesting surprise! Finally, we will discuss the velocity of a particle as compared to the velocity of waves.

4.1 Where Is a Particle?

So how can a wave be at a specific spot? What does it mean when a particle is at a specific spot? This is the case when we detect the particle at that spot. Let's call our spot (x, t) , a place in space (x) and time (t). There is only one way of knowing that a particle is at (x, t) : by arranging an interaction with it. If we do not do that, we do not know the particle is there (see Fig. 2.8). Let's think about this. How could we know a particle is at (x, t) , when we do not interact with it in any way? Suppose we produce it at one point, just left of (x, t) and we detect it at another point, just right of (x, t) . Then it surely must have been at (x, t) , halfway between, mustn't it? Well, no, since we have seen in the two-slit experiment that the particle is actually a wave that extends everywhere as long as it is undisturbed. By interacting with the particle, we disturb it. But we don't know for sure where it was before or where it will be afterwards.

So only an interaction with the particle at (x, t) is going to tell us that the particle is actually at (x, t) . When will the particle interact at (x, t) ? If we represent a particle by a wave in a field, then the amplitude of the wave at (x, t) is the field strength at (x, t) . The field strength determines the interaction strength. You can imagine that, when the field is strong, another field might feel that better than when the field is weak. But we will go into the question of interactions and how these really work later.

So, it is fair to say that the probability that the interaction takes place at (x, t) is higher when the field is stronger at (x, t) . Consequently, the interaction is most likely to take place where the particle's wave has a higher amplitude. Hence, we may interpret the amplitude of the wave at (x, t) as the probability of interacting with the particle at (x, t) . Since interacting means detecting or finding the particle at (x, t) , we see that the amplitude of the wave at (x, t) is a measure of the probability of finding the particle at that spot. To be more precise, in quantum mechanics the probability of finding a particle with wave function ψ at (x, t) is equal to $|\psi(x, t)|^2$, according to the Born rule [Ref. 1].

4.2 Waves in Space

Now imagine a single wave in one dimension that represents our particle. This wave has one wavelength, since it is a single wave. One wavelength implies one momentum. The wave has high amplitudes in many different places all over space (see Fig. 4.1), so the probability of finding our particle at some place is spread out everywhere. We can never be sure to find our particle at one particular spot. What we conclude from this picture is that when we represent a particle with one wave, we can be very sure of its momentum, but cannot say where the particle is.

Let's turn this around and see how we can represent a particle that is in one place. In that case, the particle's "wave" must look like a spike (see Fig. 4.2). The amplitude is 1 at one spot (x, t) and 0 everywhere else. So clearly, this represents a particle being at one spot. The probability of finding it at (x, t) is 1. However, this is not in itself a wave. What to do? It is important to realise that a particle does not have to be represented by a single wave. Usually, it is represented by a bunch of waves, a so-called wave packet. There can be many waves in the wave packet and, by means of superposition, they all add up to the particle's wave packet (or wave function). The wave packet consists of waves of a variety of different wavelengths

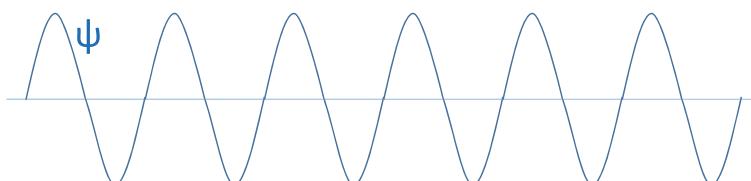


Fig. 4.1 Particle represented by a single wave: it could turn up anywhere

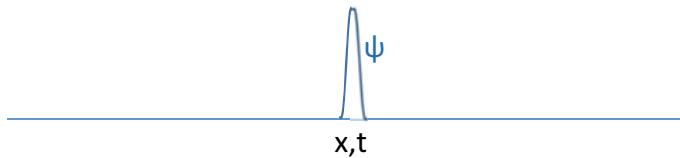


Fig. 4.2 Wave function of a particle that will most probably be found at x, t

and amplitudes. Superposition of those waves provides the amplitude distribution of the resulting wave packet. The packet shows us where the field strength is high and therefore where we can expect the highest probability of interacting with the particle.

So, what waves do we need to add up to create a spike? Suppose we take a wave that has a crest at the position of the spike. Then we add a wave that also has a crest there, but has a different wavelength and has a low where the first wave has its next crest. Clearly, these two waves add up to a high probability of finding the particle at the position of the spike (see Fig. 4.3), while the probability of finding the particle elsewhere is much lower. However, the amplitude anywhere else is not yet 0. So, let's add more waves with different wavelengths and amplitudes (Fig. 4.4), all of which have a crest at the position of the spike. It turns out that when we add enough waves (actually an endless number of waves, each with a different wavelength), we approach a spike.

Such a wave packet gives a high likelihood of finding our particle at one particular spot (x, t) . However, in order to create such a wave packet, we are required to use waves of all wavelengths. Since the wavelength represents the momentum of the particle, we see that we have no clue about the momentum of the particle. In order for a particle to be in one spot, it must be unclear what its momentum is.

What we have here is the *origin of the uncertainty relations*, which you may know from quantum mechanics. We cannot know the position and the momentum of a particle at the same time because of the fact that particles are built from waves.

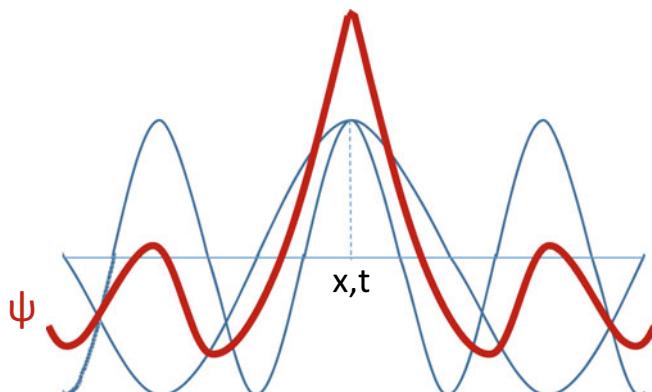


Fig. 4.3 Superposition of two waves with a maximum at (x, t)

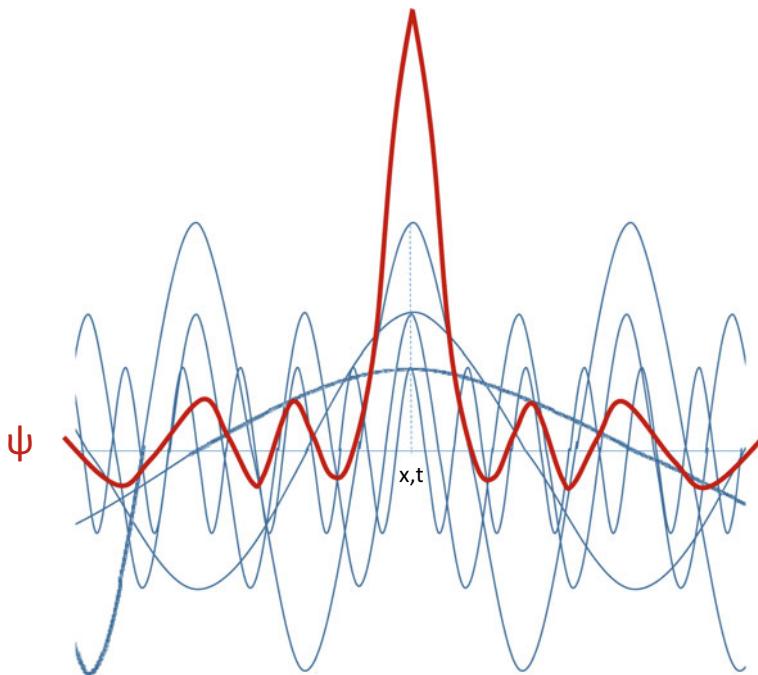


Fig. 4.4 Superposition of five waves that share a maximum at (x, t) . In order to build up the red wave, the wave that resembles the middle part best receives the highest amplitude. The more the wavelength differs from the middle part of the red wave, the lower the amplitude. This is how one can build up a spike with a certain width from a superposition of sine waves. The red wave is by no means yet a true spike, but we have only used 5 sine waves and it is already better than Fig. 4.3, using 2 sine waves. So, you may imagine that using, e.g., 100,000 waves, one could get fairly close to a spike. To get a perfect spike, one needs all wavelengths and each wave will have the same amplitude. But when the spike has some width, the amplitudes of the various waves differ. Wavelengths on the order of the width of the spike will have a high amplitude and wavelengths that deviate a lot will have low amplitudes. The wider the position “spike”, the greater the difference in amplitude between the various waves and the more important the waves become that have wavelengths resembling the width of the spike

This is expressed in the following uncertainty relation between the uncertainty in position ΔX and the uncertainty in momentum ΔP :

$$\Delta X \Delta P \geq h/4\pi \quad (4.1)$$

where h is Planck's constant.

Let's dig a little deeper into this. If the uncertainty in space is not infinite or (practically) zero, but has a finite width, e.g., like in Fig. 4.5a., what wavelengths (momenta) do we need to make such a spread in space? What wavelengths should the wave packet be made of? The math that shows us this is called Fourier analysis. This math shows us that we need a bunch of wavelengths grouped around the average

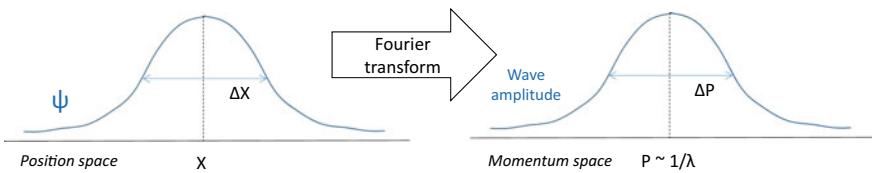


Fig. 4.5 a Bell-shaped wave function. b Bell-shaped Fourier transform of the wave function

wavelength that agrees with the momentum of the particle (Fig. 4.5b). In this picture we put wavelengths on the axis, instead of positions in space. The height of the graph shows the amplitude of the wave for each wavelength. So, we see that the further away from the average wavelength, the lower the amplitude of that wave.

The curve we see here is called a Gaussian curve. It has a bell shape. When we “Fourier transform” a bell-shaped wave function in position space (such as Fig. 4.5a), we get a bell-shaped function of all the wavelengths we need to create that wave function (such as Fig. 4.5b). We call the dimension of all wavelengths the “momentum space”. Now let’s look at the width of the bell shape. When we make the bell shape in position space wider, the bell shape in momentum space gets narrower and vice versa (Fig. 4.6). So here too we see that when we know the position of the particle with more certainty (narrower bell shape), we need more wavelengths to build the wave packet (wider bell shape in momentum space). Again, the result is the uncertainty relation between momentum and position.

In the extreme case, when the position bell shape is infinitely wide, we have just one wavelength. The critical reader will say that the wave function of one wave is an endless series of low probabilities of finding a particle at positions where the wave amplitude is 0 and high probabilities of finding it at positions where the wave amplitude is maximal. So, the distribution of the probability of finding a particle at different positions is not the same everywhere, but it does extend to infinity. It should also be remembered that the wave generally has a velocity, and the minima and maxima will change over time.

At the other extreme, when the position bell shape is a spike (the particle must be at one position in space–time), the momentum bell shape is infinitely wide. This expresses the fact that we need all wavelengths in equal amplitude to create a spike and to make sure the particle is at one position in space–time.

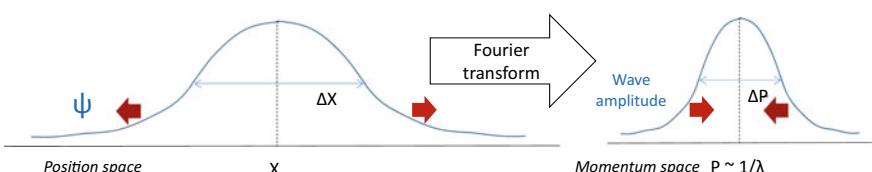


Fig. 4.6 When the wave function in position space gets wider, the Fourier transform in momentum space gets narrower

4.3 Waves in Space and the Double Slit Experiment

Let's look again at the double slit experiment and imagine an electron waving through the two slits, interfering, and being absorbed at the screen (see Fig. 2.6). The holes in the wall make sure that we know the position of the electron when it goes through the holes with a precision Δx equal to the distance between the two holes. Therefore, the wavelength of the electron going through will not be exactly clear, but will be spread out. So, the interference pattern will not be very sharp.

Now suppose we put a light bulb between the two slits. Light will be scattered by a charged particle such as an electron. This way we can identify which hole the electron went through. We set up the experiment and watch where we see a flash: hole 1 or hole 2. So, we see a lot of electrons pass through and each time we see a flash around one or other of the slits. What would be the result? As explained before, the interference pattern disappears.

So, what has happened? By identifying which hole the electron went through, we effectively improved our knowledge of where the electron is. We can be sure that it is around one of the slits. So Δx , the uncertainty in the location of the electron, has reduced to the width of the slit (or even to the location of the flash). This makes the interference pattern go away. What has happened is that the interaction (the measurement) has changed the momentum of the electron to such an extent that the bunch of waves are no longer of one wavelength. Consequently, there cannot be a clear interference pattern. The clarity in position relates to the lack of clarity in momentum (wavelength) in every step of the way.

Now let's try to lower the impact of the measurement, e.g., by dimming the light, or even by lowering the frequency of the light. Remember what happened? With less light, the flash becomes less clear. In particular, when the frequency of the light is lowered, its wavelength goes up. When the wavelength is longer than the distance between the slits, the blob of the flash widens to a size bigger than the distance between the slits and it becomes unclear which slit the electron went through. At that point, the interference pattern starts to show up again.

Before the measurement, the electron wave packet was close to a single wave. Its momentum was clear, its position was not. The single wave can show interference patterns and clear wave behaviour. The interaction with the light changed the momentum of the electron. When the momentum changes, the wave packet of the electron changes to a bunch of waves with very different wavelengths. During the interaction the momentum is unclear since it is changing. The more it changes (the stronger the interaction), the more variety in the momentum of the waves in the wave packet. Such a variety adds up to a wave packet that resembles a small peak, such as in Fig. 4.2. Hence, the position of the wave packet becomes clearer and reduces the uncertainty in the position of the electron. So, there is no escaping the uncertainty relation, which governs all waves of all particles and interactions. Effectively, we have just shown how waves can provide a position during an interaction. We will return to this concept in chap. 12, when we discuss quantum decoherence, and

we will dive deeper into the way the wave changes through interactions in order to measure the position of a “particle”.

One could argue that the two slits themselves constitute an interaction: they reduce the location of the particle-wave to two locations (the two slits). This interaction also makes the momentum uncertain. Before going through the slits, the wave packet has a certain width. When going through the slits, the wave packet gets reduced in width by the number of slits (two) and the width of each slit. So, in momentum space, the wave packet has to become wider (i.e., there will be more wavelengths present in the wave packet).

However, when the wave gets a chance to progress undisturbed after the slits, something else happens. When there are a variety of momenta available in the wave packet, it turns out that these waves start to spread. When they spread, they interfere with each other differently. That interference reduces the influence of waves with a wavelength that deviates a lot from the average. Also, the interference makes the wave packet spread in position space. Mathematically, this spread is directly proportional to the time elapsed.

So, when a wave packet is allowed to travel undisturbed, it gets wider in position space and narrower in momentum space (see Fig. 4.7). You can also view it this way: when the wave is undisturbed for some time it becomes unclear where the particle is, and so it becomes clearer what its momentum is. This means that when the screen is moved to a greater distance, the interference pattern will become clearer!

We can conclude that it requires an interaction to improve the clarity in the position of a particle, while necessarily making the momentum less clear. Equally, we can say that when left undisturbed, the clarity in the position is reduced while the momentum becomes clearer. And so again, there is no escaping the uncertainty relation. The particle behaviour comes out in interactions, while the wave character comes out when the wave packet is left undisturbed. The uncertainty relation is not magic, but is simply a consequence of the way the momentum and position characteristics of

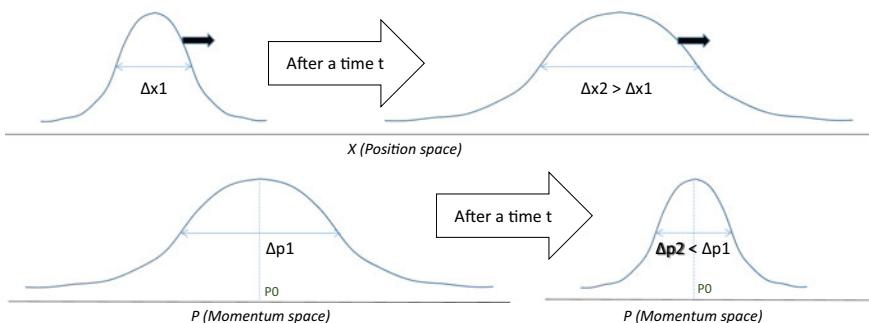


Fig. 4.7 Spread over time in the wave packet. In the top figure, the wave packet moves to the right in space with the average velocity of the particle. The width of the packet in space increases. In the picture below, in momentum space, the wave packet does not move (since the average momentum stays the same at P_0), but in time it does narrow as the packet in position space widens

waves are related. The fact that we all obey the uncertainty relations is a consequence of the fact that we are all made of waves.

4.4 Waves in Time

So far, we have been discussing the wave function in position space, but it changes in time as well. We saw before that when we paint one point on a rope red, the red dot will go up and down at a particular frequency whenever a wave passes by. A frequency is nothing else than a wave on a timescale. This wave also has a “wavelength”. In the dimension of time, the wavelength is called the period T of the wave. The frequency is equal to $1/T$. Clearly, when the period gets shorter (the red dot goes up and down in a shorter amount of time), the frequency is higher.

What happens when the wave lasts forever, i.e., when the red dot keeps going up and down forever? In this case, the frequency of the wave never changes. We can represent this with a wave in time with one particular frequency. Since frequency \sim energy, the energy does not change and is perfectly well defined. In this case, the lifespan of the particle has no limit, so we cannot define a period in time during which the particle exists and before/after which it does not. It simply lives forever, meaning that the wave keeps going forever.

What happens if the particle exists only during a very short time? Such particles are often seen in particle physics. We will meet some examples later, such as virtual particles or resonances. How can we create a wave in time that corresponds to a small peak in time during which we may find the particle alive? This means that the particle wave must have an amplitude during that short period of time and otherwise be zero. This looks much like the spike we saw earlier in position space. In order to create such a wave, we are going to need a superposition of a whole bunch of waves of different frequencies (see Fig. 4.8).

So, similarly to the case of position and momentum, we conclude that there is an uncertainty relation between time and energy:

$$\Delta t \Delta E \geq h/4\pi$$

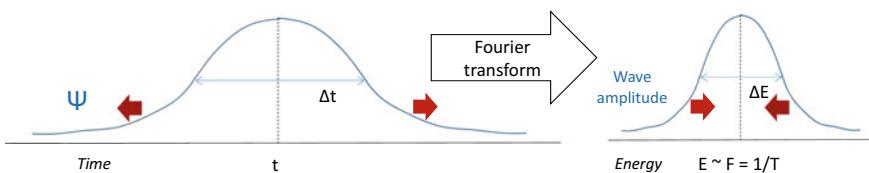


Fig. 4.8 The longer a particle lives (the wider Δt), the smaller the range of frequencies needed to create the corresponding wave function. So, the smaller ΔE

Again, it is the wave character of matter that makes it impossible to know exactly the energy of a particle at a specific time, or to be certain of the energy of a particle during a time Δt . This property is relevant for virtual particles: particles of energy ΔE that can exist during a maximum time Δt .

How does that work? What actually happens is that a virtual particle consists of a bunch of waves that add up to an amplitude during a time Δt , but that cancel each other out after that time. We will come back to this later when we discuss virtual particles in detail. But to do this, we first need a much better understanding of the behaviour of fields and waves!

So, if you ever read about the uncertainty relations before and wondered why energy is related to time (and momentum to space), it is because energy is proportional to the frequency of a wave and a frequency is related to a time period. And likewise, momentum is related to the wavelength which is a stretch of space.

4.5 A Particle Is a Bunch of Waves

An idealized real particle is a particle that consists of one wave, with one wavelength (one clearly defined momentum) and one frequency (one clearly defined energy). However, such a particle does not exist. Take for instance a photon that comes from a galaxy far, far away and hits the detector in a powerful telescope. The photon may have lived for billions of years, but it is not infinite! Such a photon would come very close to the idealized particle with a clear momentum and energy but not exactly!

Consequently, the difference between virtual particles and real particles becomes a little shady. The shorter the particle lives, the more uncertain its energy, and the shorter the distance it travels, the more uncertain its momentum. Until finally its time and space dimensions are small enough for it to be a virtual particle. You may have read elsewhere that virtual particles are responsible for the forces we experience. This looks like an important difference between the two types of particles, but actually, a real photon can exert the same force, e.g., on an electron as a virtual photon does. Both carry energy and momentum from one charged particle to another. So, the difference between virtual and real photons is not straightforward. We will come back to this in Chap. 10.

There is another thing that the uncertainty relations tell us. Take the energy of a particle to be $E = MC^2$. Now let's limit the uncertainty ΔE in the energy of this particle to be equal to that energy MC^2 . Then we find:

$$\Delta t \geq h/4\pi MC^2$$

So, a particle of energy MC^2 cannot be located better in time than this value. It must be smeared out in time! One can also define a length called the Compton length. It is considered to be a measure of the size of a particle of mass M . It is defined as

$$\lambda_c = h/2\pi MC$$

When we take MC to be a measure of momentum, the Compton length is basically a measure of the wavelength of a particle at rest. Consequently, we get for the time Δt :

$$\Delta t \geq \lambda^c / 2C$$

Both dimensions λ^c and Δt are built up from Planck's constant h , the rest mass M of the particle, and C . All three are absolute values and combine relativity with quantum mechanics. So, the basic properties of the vacuum and the fundamental constants of nature tell us that, in a wavy world, particles are smeared out in space and time.

What we conclude from this discussion is that all particles are built from a bunch of waves of different wavelengths and frequencies: a wave packet. When the particle gets more limited in space and time, the number of waves needed to describe its wave function increases, as does the range of wavelengths and frequencies involved in these waves. In general, the wave packet is smeared out in space and time which means that we will never be able to pinpoint a particle at an exact location in space–time.

4.6 Velocity of Particles and Waves

Now let's look at the velocity of particles. Is that equal to the velocity of the wave? Not exactly. Now that we have concluded that a particle consists of a wave packet, a bunch of waves, we need to take a look at what is called the group velocity of the bunch. Why is this any different?

As an example, we look at a “bunch” of two waves of slightly different wavelength and frequency (Fig. 4.9). They move at slightly different velocities. The result is that

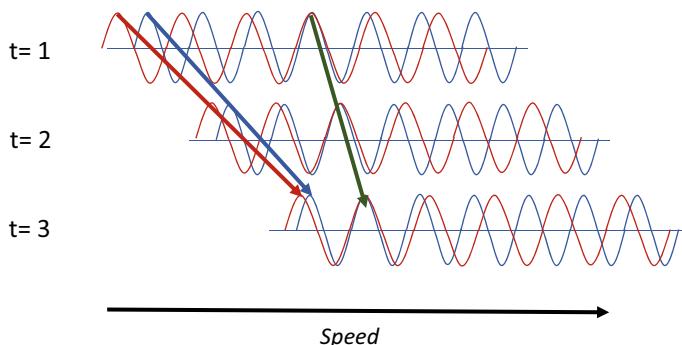


Fig. 4.9 The blue wave and the red wave create a superposition. The blue wave moves to the right at a certain velocity indicated by the blue arrow. The red wave moves just a little faster to the right, indicated by the red arrow. The top of the superposition moves at a much lower velocity to the right, as indicated by the green arrow

the top of the bunch (where the two waves create a maximum because they are both at the top) moves at a different velocity than the individual waves. In this particular case, the group velocity is half the velocity of the individual waves. From the picture it becomes clear that this also depends on the difference in wavelength and frequency. It turns out that, when the spread in wavelength and frequency becomes smaller, the group velocity gets closer to the velocity of the individual waves. When we have just one wave, the “group velocity” is equal to the velocity of the wave.

The fact that the velocity of a wave can be dependent on its frequency is called “dispersion”. The velocity of the individual waves (also called the phase velocity) can be spread out. When the group velocity approaches the speed of light, the phase velocities of the individual waves get closer together. At the speed of light (massless particles) the group velocity is equal to the phase velocity.

The group velocity can be found by looking at how the frequency depends on the wavelength. The equation that describes how the frequency depends on the wavelength is called a dispersion relation. When this relation tells us that the frequency is a constant divided by the wavelength (i.e., $F = C/\lambda$), the group velocity is equal to the phase velocity (as above with light). However, *when the frequency depends in a different way on the wavelength, the velocity becomes dependent on the frequency*. We will see some examples of this later, e.g., when we treat waves with mass.

Sometimes it is important to distinguish between group and phase velocity, as their behaviour can be very different. When the group velocity gets close to 0, the phase velocity can in principle exceed the speed of light. This is a strange situation that we will not explore further (see, e.g., [Ref. 10]), but it does not allow any information or energy to be transported faster than light.

So, a wave front or group cannot go faster than C . Put differently, a signal cannot move faster than C . The way this works is that, when the faster wave reaches the wave front/group, its amplitude starts to diminish. The slower waves experience the same thing when they appear at the tail of the wave front/group.

Although this process maintains the group as a sort of bump propagating through the medium (at max C), the bump will get distorted in time as well. It will start to broaden as slower waves extend the tail and faster waves extend the front. This can happen when the velocity of the bump is lower than the maximum velocity in the medium, since only then we can have faster waves extending beyond the bump. The slower the bump moves, the more spread out it can become.

After enough time, the bump broadens as a result of this process. Consequently, when a bump starts off as a sharp peak (we know where it is, i.e., where the particle is), it will broaden after a certain time and we will no longer know where exactly it is. When the bump is still sharp, it must be built from waves with very different wavelengths (and velocities!) and this will make the bump broaden even more. After some time, when the bump broadens, it will be made from waves that are much less diverse in wavelength (and velocity!) and the particle will have a sharper momentum.

In most cases, the group velocity is also the velocity at which energy or information is conveyed along a wave. Hence, the group velocity can also be considered as the signal velocity. When waves move through an absorbing medium, this does not generally hold, as energy gets absorbed.

A common way to extend the concept of group velocity to complicated media is to consider a (spatially damped) plane wave solution inside the medium. We will do the same and base the rest of our story on plane waves moving through media. In most cases when we talk about wave velocity, we will mean group velocity and we will view the wave front/group as a plane wave. This will make things easier to understand. Moreover, since the group or bump represents a particle, the group velocity is the most relevant for understanding how the world is built from waves.

Chapter 5

The Potential of a Field's Elasticity



What is the origin of a wave? Why is something waving in the first place? To understand that we need to look at the concept of potential energy and how that gets exchanged with kinetic energy. Just like the pendulum of a classic clock. The pendulum determines the pace of the clock. It does so by swinging from left to right at a fixed frequency. The swinging motion is characterized by a continuous exchange of potential energy and kinetic energy. It has maximum potential energy when it is at the top and has a maximal kinetic energy when it is at the bottom (see Fig. 5.1)

This is an example of a harmonic oscillator. The main characteristic of a harmonic oscillator is that it has a periodic movement that can be described by a wave and it has a typical frequency that depends on the characteristics of the object that is oscillating. That frequency does not depend on the amplitude of the wave or on its wavelength or on the way the wave was triggered. An example is a tuning fork. You can hit it softly or hard, but the frequency of the sound wave it produces is always the same.

Let's see if we can compare such an oscillator to a wave.

5.1 Exchanging Energy in a Field

Let's look at a wave in water. The crest of a wave in water consists of water molecules that are situated above the average water level. Consequently, they have a potential energy. When the wave moves to the right, the water molecules start to drop, thus getting kinetic energy. This kinetic energy forces them to continue dropping until below the average water level. At that moment they are in a trough of the wave. So apparently a water wave is nothing but a continuous transfer of potential energy into kinetic energy and back. Just like the pendulum.

The same works for a rope in space (in the absence of gravity). When the left end of the rope is moved up and down, the up position pulls up the parts of the rope directly to the right as well. These other parts pull up the parts to their right and so on. In the meantime, the left end is pulled down again and the left end of the rope

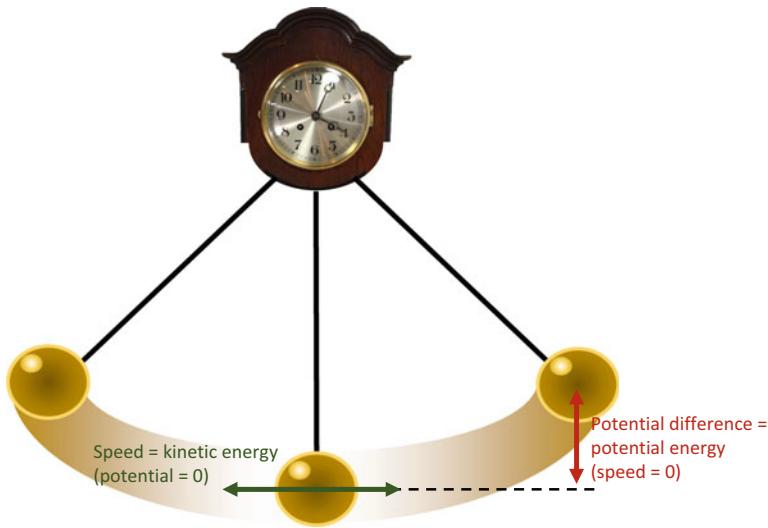


Fig. 5.1 The pendulum as harmonic oscillator. It oscillates between the two extreme positions and between potential energy at both ends and kinetic energy in the middle. The frequency of oscillation primarily depends on the length of the pendulum. For small amplitudes, it does not depend on the amplitude. Hence, it will not depend on the initial kick that got it going.

now starts to pull down parts of the rope to its right. So here we see that the potential energy of each part of the rope is the result of the parts next to it pulling it up or down. We get the same interplay between potential energy and kinetic energy, although the cause of the potential energy is not gravity, but the elasticity of the rope.

When we describe fields, we must assume that there is a kind of elasticity in the field in order for waves to be able to move across the field. There must be some cohesion that enables one part of the field to transfer energy to another part of the field.

The elasticity of the field or rope determines how strongly one part of the rope is connected to the next part. When this is strong (the rope is not elastic), the next part of the rope will experience a strong potential when there is a difference in amplitude between the two parts of the rope. That potential, in turn, determines how fast the next part of the rope will respond to the difference in amplitude (see Fig. 5.2). Basically, pulling the rope up will mean that the part immediately next to it is pulled up too. You can imagine that the stronger that potential, the faster the reaction. Hence, the pulling up and down of such a rope will be propagated through the rope very quickly: the velocity of the wave will thus be high. So, the speed of waves in a rope depends on the elasticity of the rope. In addition, the wavelength will be long.

On the other hand, when the elasticity is high, we have to pull the rope a lot before the next parts begin to move. The velocity of a wave in such a rope will be low, while the wavelength is short.

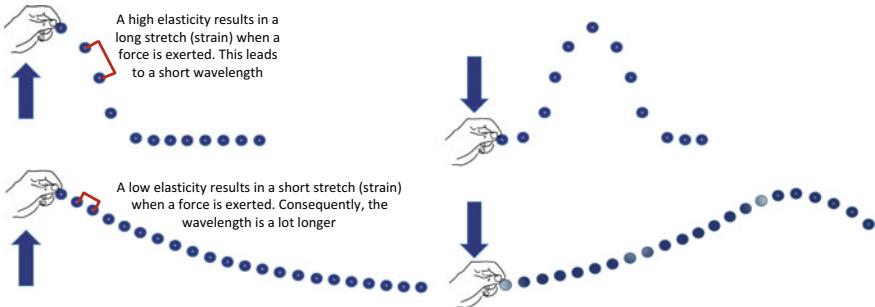


Fig. 5.2 Elasticity determines the relation between frequency and wavelength

Comparing the rope to a field, we see that the elasticity of a field determines the velocity of waves, and hence the relation between frequency and wavelength. Earlier we saw that this is expressed in the formula $F = V/\lambda$.

For all massless fields in nature, we know that their waves have exactly the same velocity: C , the speed of light. So, we may conclude that all massless fields have the same elasticity. This equivalence is very strong as we saw in the measurement of gravitational waves in 2017, coming from a staggering distance of 130 million light-years away. The event provided a limit on the difference between the velocity of electromagnetic waves (light) and that of gravitational waves. Assuming the first photons were emitted between zero and ten seconds after the peak gravitational wave emission, the difference between the velocities of gravitational and electromagnetic waves is constrained to lie between -3×10^{-15} and $+7 \times 10^{-16}$ times the velocity of light [Ref. 53, 54]. So, this is a very small difference, indicating that the elasticity of massless fields is indeed very much the same.

Why should this be so? Why would each field not have a different elasticity? It is too coincidental that this velocity is the same for all fields. Therefore, one could suggest that this elasticity is determined by the nature of space–time itself. In any case one can conclude that the elasticity of massless fields is a shared characteristic. What about massive fields? They all seem to share the characteristic that the speed of light is a maximum speed. Suppose they were massless. Then we could expect their speed to be C as well and they would share the same elasticity as other massless fields.

From such a shared characteristic, it is reasonable to assume that there is only one carrier of all these fields, which is the vacuum. Just as the atmosphere that can carry all sorts of fields, the vacuum is one kind of stuff in which everything is waving. What that stuff is we don't know, but it seems to have one type of elasticity and it carries the characteristics of space–time. There are some suggestions that the vacuum could be made of “information”. So, a smallest bit of space–time would contain one bit of information. However, every bit of information also needs a carrier, so what is the carrier of that information? None of this is proven, verifiable, or falsifiable, though.

The elasticity of the vacuum sets the maximum velocity of waves at C , but we saw earlier that the phase velocity of waves that make up a group or bump in the

field could be greater. How can that be? This too is a characteristic of the vacuum. The elasticity of the vacuum does not prohibit individual waves from going faster when a bump is made of different waves. As we will see later this is possible for waves that experience an altered elasticity, in particular when a particle has mass. Massless particles are tied to the elasticity of the vacuum and travel at C. Massive ones must travel more slowly as a group, but their individual waves can go faster, although they diminish beyond the wave front. When such a bump approaches the speed of light, the difference in velocity must diminish, and as we will see later, its dispersion relation starts to behave more like that of a massless particle.

You can compare the electromagnetic field to oscillations in a large block of rubber. If you do [ref. 38], you find that they behave exactly alike. The velocity of waves is determined by the elasticity of the block of rubber, and hence for the vacuum this velocity is determined by the elasticity of the vacuum. The Michelson-Morley experiment made use of the wave characteristics of electromagnetic waves. Their measurements showed that the light always has the same velocity. In fact, they showed that the elasticity of the vacuum is the same independently of how one moves through it. One might say that in that sense they actually proved the existence of the vacuum-ether, except that this ether has different characteristics.

Now an interesting question comes to mind: what would happen if we changed the elasticity? To investigate this, let's first take a look at waves in different media.

5.2 Waves in a Medium

Consider what happens to waves when they enter a different medium [ref. 10]. For instance, when a wave in water of a certain depth propagates into shallower water, the velocity and wavelength change (see Fig. 5.3). We saw this before when we were talking about waves closing in on the beach.

When light enters glass, it interacts with the medium. The result of that interaction is that its wavelength shortens, much as happens for the water waves in Fig. 5.3. So, entering a different medium changes the relation of a wave between its frequency and its wavelength. In optics this principle is used to create lenses, for instance. Each different medium has a property that is called its refractive index (N). This property describes, e.g., for glass how much the wavelength of light is lowered. It is defined as

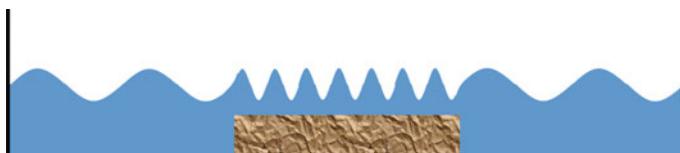


Fig. 5.3 Both wavelength and velocity decrease when water waves enter shallower water, and both increase when they enter deeper water

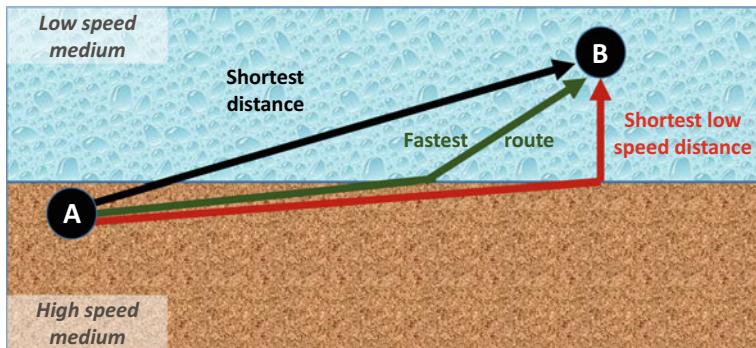


Fig. 5.4 The fastest route in two different media

$$N = \lambda_0 / \lambda$$

where λ = wavelength of light in the glass and λ_0 = wavelength of the same light in vacuum. So, the wavelength shortens, but the light does not get extra energy when it enters the medium! Consequently, its frequency must remain the same. Since $F = V/\lambda$, this means that when the wavelength gets shorter, the velocity must also be lowered by the same amount.

There is an interesting principle in optics, called Fermat's principle. It states that the path that light takes between two points is the path that takes the least time for the light to cross. In vacuum it means that light travels in a straight line, but what happens when light enters a different medium? In that medium it travels more slowly. Take for example the situation shown in Fig. 5.4.

Let's first take an example from daily life. Suppose you are on the beach standing at point A in Fig. 5.4. Someone is drowning at point B, so you have to get there as fast as possible. Running on the beach is faster than swimming in water, so what would you do? Would you run on the beach until you have the shortest distance to swim? Would you ignore this velocity business and go in a straight line? A straight line implies a lot of swimming, so it is certainly not the fastest. Taking the shortest swim could be faster, but you do have to cover some extra distance. It turns out that the fastest is to run more beach than a straight line, but less distance than you would have to have the shortest swim. How much less distance? That depends on the difference in velocity between running and swimming. If you happen to be the world's fastest swimmer but you are no good at running and you swim as fast as you run, you would go in a straight line. If you run 100 times faster than you swim, the fastest will be close to the path with the shortest swim. So, it turns out that the velocity-ratio between swimming and running determines the route you would take.

The same goes for light. Fermat's principle says that light seeks the shortest route between A and B, and what this route must be depends on the relative difference in velocity. That difference directly depends on the difference in wavelength. And the difference in wavelength is determined by the refractive index.



Fig. 5.5 Applications of the change in velocity and wavelength in a lens and a prism

So, the refractive index determines the route! Now we see why it is called the refractive index. The higher N is, the bigger the difference in velocity, and the more “bent” the path that the light will take when it enters the other medium. So, the path is more “refracted” when N is big. Hence, the name “refractive index”. An application of this can be found in a lens (see Fig. 5.5).

The refractive index of materials such as glass or Perspex depends on the wavelength of the light. Hence, N is different for different colours (wavelengths) of light. The dependence of N on the wavelength is called dispersion. A prism uses this effect to split white sunlight into all the colours of the rainbow.

More interesting is this question: how does the light know what medium lies ahead of it? When you are on the beach you oversee the situation and make a choice, but you wouldn't tell me that light does that, would you? Let's keep this a mystery for a while, but we will come back to it in Sect. 9.3 when we discuss the path integral. That's when we will resolve it. But you will need to know more before we can do that.

Another question is why the light goes slower in a medium. We have seen that the speed of light depends on the elasticity of the vacuum. A medium is also something that exists in the vacuum, e.g., glass is just a bunch of atoms that reside in the same old vacuum. Nothing more. So why would the elasticity of the vacuum change all of a sudden?

Let's go a little deeper. The elasticity was determined by the potential between the adjacent parts of the field. We do not know what the vacuum is made of, but we can assume that the elasticity can be described by such a potential. In the glass we find that the light does not only experience the potential of the vacuum that determines its speed, but also another potential: the light interacts with the electrons of the glass atoms. One can view this as an extra potential that changes the elasticity experienced by the light.

This effect has profound consequences for the behaviour of fields. In fact, the behaviour of waves in fields is determined by all the potentials they experience. In the mathematical description of quantum field theory, such potentials are gathered together in what is called the Lagrangian. That formula is in turn used to understand the kinematics of the waves. One may say that all the potentials together determine how the waves move in the field. That makes sense. We do the same when we look at a falling object: the potential created by gravity determines how fast the

object will gain speed. On earth this is faster than on the moon, since on earth the gravitational potential is higher. So, we use our knowledge of the gravitational potential to understand how the object is moving.

In the next chapter we will go more deeply into this effect. We will look at how an extra potential can change the elasticity that is experienced by waves. And we will look at some of the consequences. We will use this effect to explain the relativistic behaviour of objects. In later chapters we will make this view more precise as we go along.

Chapter 6

A Wave of Relativity



The theory of special relativity is one of the pillars of quantum field theory. The origin of relativity lies in the solution of the ether problem as described by the Michelson-Morley experiment. We saw earlier that the experiment made clear that the speed of light is the same no matter how fast you go. This is considered to be a “law of nature” and we related it to the elasticity of the vacuum. H.A. Lorentz wrote down the formulas to understand how velocities must be added in order to have a constant speed of light. However, it was Einstein who explained why that must be so [Ref. 42]. He stated that laws of nature should be the same in each frame of reference, i.e., no matter what relative velocity you have. And for him, the constancy of the speed of light was a law of nature.

The consequence of these observations and notions is that we get some strange behaviour we are not familiar with in everyday life, largely because the speed of light is much faster than anything we usually encounter. So, let's take a look at some of these effects.

6.1 Wave Velocity

So far, we have found that space–time is not empty. It is filled with a medium we call the vacuum. We, our world, particles, and anything else are just waves in this medium. The vacuum has a particular elasticity, which gives rise to the velocity of the waves as determined by that elasticity. This velocity is C , the speed of light (and of all massless waves).

As we will see shortly, there is a reason why massive waves move at velocities lower than C . But nothing goes faster. So, the elasticity of the vacuum governs how fast we can move. Faster than C is not possible: the (group) waves simply cannot move faster than the elasticity allows them to.

When a massive wave moves almost as fast as the speed of light, we can still keep adding energy to it. However, this will no longer translate into a substantially higher

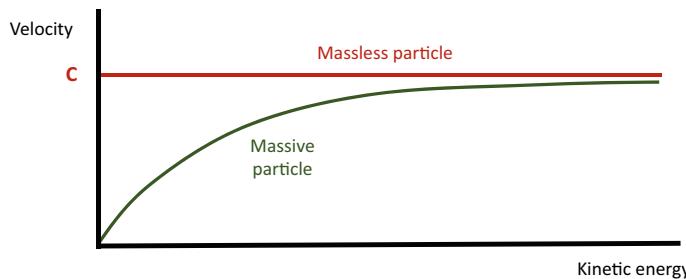


Fig. 6.1 The speed of a massless wave such as a photon is always C, independently of its energy. A massive wave, on the other hand, will start at speed 0 when the energy is 0. It will gain speed with any extra energy it gets. With enough energy that speed will get close to C, but will never reach it

velocity. Still, it must translate into higher kinetic energy and higher momentum. So, the relation between momentum and velocity cannot be linear for anything that has mass (see Fig. 6.1)!

In order to explain how this can be, we start in a rather unusual manner: by looking at the behaviour of mass and energy. Then we will look at other effects such as the metric of space–time, and effects such as length contraction and time dilation.

6.2 How Does a Wave Become Massive?

So, we see some strange behaviour in relativity. While massless waves always have the same velocity C, massive waves do not. Where does this difference come from? In fact, what *is* mass in our wavy world? And why do massive waves behave differently compared to massless ones?

6.2.1 A Game with Rope and Springs

Imagine a wave in a rope. You have to move the end of the rope up and down in order to produce a wave. The elasticity of the rope determines the velocity of the wave and with it the relation between frequency and wavelength.

Imagine such a wave in a rope connected to springs that are attached to the floor (see Fig. 6.2). What does this do to the properties of the rope? First of all, before you attached the springs it was easy to create a wave. With the springs attached you have to put a lot more effort in to get a wave going. In essence, it takes a higher energy, hence frequency, to start a wave.

The second thing that happens is that the wave in the rope has to pull the next spring up. So, you not only have to pull the rope itself up, but also the spring that is attached. As a result, there is an extra force between your hand and the spring

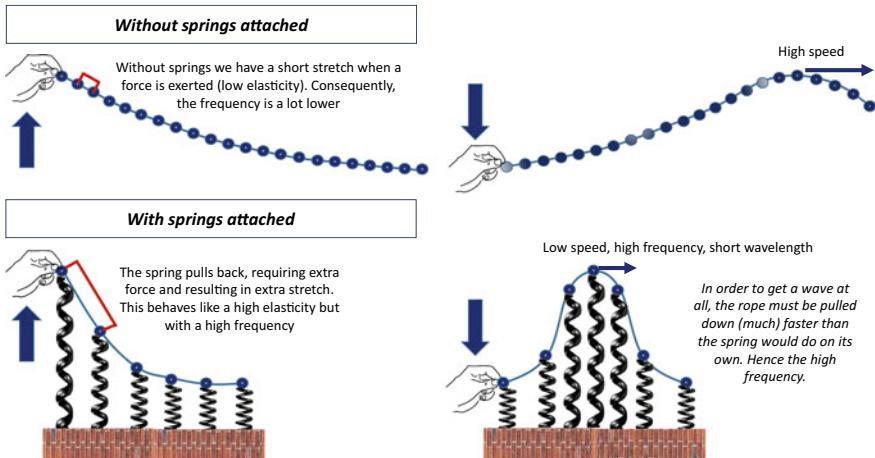


Fig. 6.2 A rope attached to springs requires more energy to start a wave and increases its elasticity, hence lowering the velocity of the wave passing through, while raising its frequency

pulling in the other direction. The result is that the rope gets stretched more. It gets stretched more because there is an extra force between one part of the rope (+ spring) and the next part of the rope (+spring). Stretching the rope takes time. That means that a signal will transport through the rope at a slower pace. Consequently, we can view a rope that gets stretched more as having a higher elasticity. A rope with a higher elasticity produces a wave with a lower velocity and shorter wavelength (see Fig. 6.2).

The third thing that happens is that energy gets stored in the springs. When you stretch a spring with your hands, you notice that it takes energy to do that. When you release the spring, it contracts. The energy that you stored in the spring is released as well. So, any stretched spring contains energy. In Fig. 6.2 you can see that a wave in a rope with springs attached makes a number of springs stretch. So, a wave in such a rope stretches a number of springs at any time during its existence. The sum of the stretch in these springs is equal to the energy stored in them. So, getting a wave going in such a rope would imply that some of the energy of the wave gets stored in the springs.

We see that attaching springs to the rope changes the properties of the medium we call “rope”. In reality we do not have springs and the vacuum is not a rope. So, what could play the role of the springs? In essence, a spring just creates a potential difference when it gets stretched. So, we could say that any potential difference working on the rope would act like a set of springs attached to the rope. Such a potential difference would show a different potential at the wave crest compared to the wave trough.

What could cause such a potential difference? Basically, any field that interacts with the medium (the rope) could do that. One example is the Higgs field. This field is all around (like all fields), but it differs from other fields in that it has a

non-zero value everywhere. Consequently, any field interacting with the Higgs field experiences an extra potential difference. This changes the properties of the vacuum for that field, as if the vacuum was attached to springs for this particular field. The change in properties is always that it increases the elasticity, resulting in a lower velocity, shorter wavelength, and higher frequency.

You may consider the field with springs as a different medium compared to a field without springs. You can compare it with light in a medium as discussed in the previous section: there are many similar effects. The light interacts with the electrons, e.g., in glass, and as a result feels a potential. This slows the light down and reduces its wavelength. You may state that a reduced wavelength should imply that the momentum goes up. However, the wave also gets slowed down. Let's compare this to massive particles. Their momentum is determined by their speed and mass. When we add a potential difference to a massless wave, we make it “massive” and it starts to behave like a massive particle. As for massive particles, in order to keep the momentum the same, its speed must be reduced when its mass increases. Either way, adding a potential to a wave reduces its wavelength as well as its speed, and the momentum stays the same. As long as the potential stays the same, the momentum bears an inverse relation to the wavelength ($P \sim 1/\lambda$), but when the potential changes, that relationship changes too.

Let's examine the consequences of the rope + springs model for massive waves.

6.2.2 Consequence 1: You Cannot Go Faster Than Light

The (group) velocity of waves in a field with springs attached will not be able to exceed the speed of light. Their velocity must be lower, no matter how much energy is put in. This is caused by the fact that each potential makes the wave more elastic. Hence, the (group) velocity of a massive particle wave will always be lower than that of a massless wave.

6.2.3 Consequence 2: The Relation Between Frequency and Wavelength Depends on the Mass

The relation between frequency and wavelength is different for massive waves compared to the relation in vacuum, without the “springs”, as it were. This difference is indeed caused by the “springs”. In their presence, the total energy in the wave is now a combination of the energy in the rope (the moving wave) and the energy in the “springs”, as if there were two frequencies combined in one wave: $f_1 + f_2$.

We cannot just add these frequencies. The rope and the springs have to be considered separately, as if they were two different dimensions. Let's see why.

First, imagine what happens when we put energy into the springs alone: if we could just trigger one spring, it would not change the velocity of its neighbour, as it is not connected to it when there is no rope. So, we could say that the elasticity does not exist and no wave and no velocity results from that. It cannot propagate any momentum because it is not connected to its neighbour.

So, the only way for these springs to propagate momentum is by being connected to the rope. The rope will do the propagating. But before we can get any sort of wave, we must first put in enough energy to overcome the pull of the spring. The spring constant determines the wave (particle) to be created. Put differently, a wave can be created when the spring is excited to the point where it can hold at least one wavelength of the particle. This wavelength is related to the spring, and not to the wave propagating along the rope plus springs. The spring is a harmonic oscillator and the way it oscillates has a wavelength and a frequency of its own. The wave that propagates momentum is a wave in the rope, across the springs (see Fig. 6.3).

So, we must conclude that we are dealing with two different types of wave in this system. Before we can set up a wave in the rope, we must first put in enough energy to excite the spring. We will not be able to get any sort of wave going in the rope, before that condition is fulfilled.

After we have done that, we can add more energy and this will flow into the rope to get a wave going. When this is just a tiny bit more energy than the energy put in the springs, the wave will go at a velocity entirely determined by the increased elasticity caused by the springs. At this point we can say that the elasticity is dominated by the springs. The stronger the springs, the higher the elasticity of the rope and the lower the velocity of the particle at a particular, low energy.

Now suppose we identify the strength of the springs (the spring constant) with the mass of the particle. When we say that we first have to create an oscillation in the

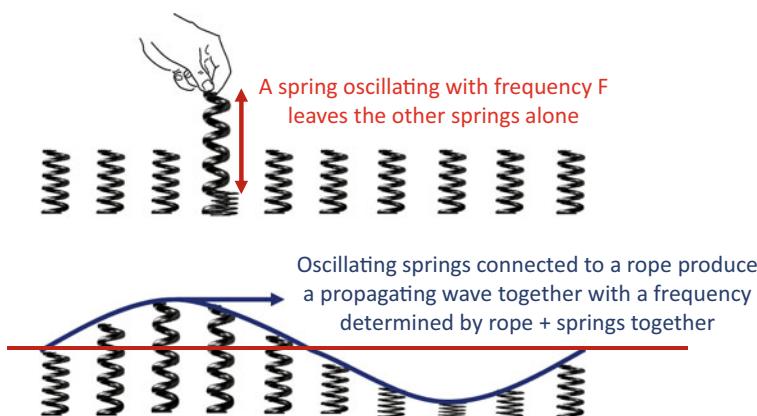


Fig. 6.3 The separate frequencies of the spring and the rope. The spring causes a potential difference. When it is pulled and released, it will vibrate at a frequency that is determined by the properties of the spring alone. But the rope inputs its own characteristics and together they determine how a wave propagates along the rope + springs

spring, what we are really saying is that we first have to create a “particle”, or better: a field quantum (see Chap. 7 on field quantization). This is a lump-sized wave in the “rope field”, but its mass is determined by the springs. Stronger springs (higher spring constants) then correspond to a bigger particle mass: more energy needs to be put in to excite the springs.

When we put in a little extra energy ΔE , the wave gets going, but its velocity depends heavily on the elasticity. Higher mass = stronger springs = higher elasticity = lower velocity when the same ΔE is added. So, the bigger the mass, the lower the velocity gained with a little extra energy. This is called inertia: to move a bigger mass you have to put more energy in.

Now let's look at the situation where we put a lot of extra energy in, so we get a serious wave going. Suppose there is so much excess energy that the amount of energy going into the wave dominates the energy that went into the springs. The frequency that the kinetic energy represents is much higher than the springs want to oscillate at. So, the rope is vibrating faster than the springs can follow. Consequently, the wave behaves almost as if the springs are not there: they do not take part in the wave movement as much as they would otherwise have.

Now you can imagine that the behaviour of the system is determined more (but never entirely) by the rope and the elasticity of the rope itself (without springs). You can also think about the extreme case of very weak springs: the strength of the springs would be negligible compared to the energy in the wave. In this case, the velocity of the wave would be (almost) completely determined by the elasticity of the rope. Going back from the rope to the vacuum, this velocity will be determined by the elasticity of the vacuum. We know that velocity: it is C.

So, the velocity of the wave varies between very low and almost C. This velocity is entirely determined by the relation between the energy put in the springs and the energy put in the wave. The velocity is low when the energy is so low that the behaviour of the system is predominantly determined by the springs (the Higgs potential), and the velocity is almost C when the energy is so high that the behaviour of the system is predominantly determined by the characteristics of the rope (the vacuum).

We expressed this relation in Fig. 6.4. For there to be any sort of wave, the minimum energy required is on the vertical axis. The excess energy does not go into the springs (as this is quantized to one field quantum), but all goes into the momentum, which is determined by the rope.

The picture shows how we can see the two different types of energy as two dimensions. When we combine two different dimensions, we have to add them using Pythagoras.

So, we get the formula:

$$E^2(\text{total}) = E^2(\text{rope}) + E^2(\text{springs})$$

Or, when we realize that the frequency of the wave is related to the energy, i.e., $F \sim E$,

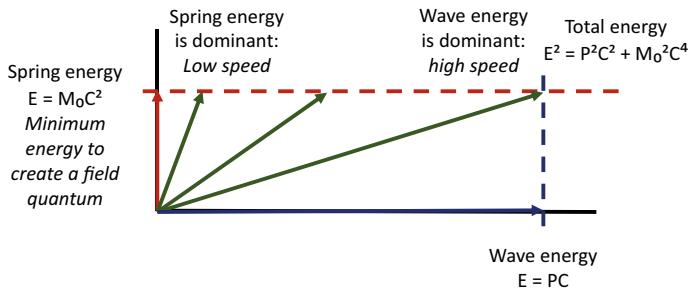


Fig. 6.4 The relation between energy in the springs (potential difference) versus energy in the rope (vacuum)

$$F^2(\text{total}) = F^2(\text{rope}) + F^2(\text{springs})$$

For the rope alone we have $E = PC \sim C/\lambda$. For the springs we introduce the standard relativistic formula for the energy of a mass at rest $E = M_0 C^2$. Then we have

$$F^2(\text{total}) \sim E^2(\text{total}) = P^2 C^2 + M_0^2 C^4$$

This is the relativistic formula for the total energy of a system. It is in fact called the “dispersion relation” for massive particles! So, we see that for massive particles the dispersion relation is different than for massless particles: in the equation $F \sim E = PC$, the phase velocity is always C and does not depend on the frequency (hence massless waves are dispersionless). In the case of massive waves, it differs depending on the frequency. A particle is made of a group of waves, as we saw before, and these waves all have (slightly) different frequencies, in accordance with the energy uncertainty relation. Now we see why these waves will also have a different velocity! As we saw before, the consequence is that the bump that represents the particle will have to get broader when the different waves have different velocities. So, the bump will disperse. It will get broader and reduce in height (amplitude). Hence, the “dispersion relation” [Ref. 10, 11].

We assumed that the energy of the springs equals $M_0 C^2$ and we found that we can explain the relativistic dispersion formula for total energy using the picture of a rope with springs attached. In the next section we will see how that impacts the group velocity of the waves in such a rope with springs.

6.2.4 Consequence 3: Mass = Inertia

In our daily experience we know that mass has inertia. You need to put energy in to make the velocity go up. In the discussion so far, we have already seen that, when the mass is higher, we need more energy to make the velocity go up.

Something very massive has a lot of strong springs attached, leading to a very high elasticity. Getting a wave going in such a “rope” requires a lot of energy. So does speeding that wave up. Just imagine a very high elasticity and a wave going through. When you want to speed up such a wave you must give a lot of extra swing to the rope. A lot of extra swing is a lot of extra energy, especially with strong springs attached.

Concluding, we can say that how fast the velocity goes up (the acceleration) is determined by the ratio between the amount of energy you put in (to increase the speed) and the total mass energy of the wave. Since the mass is determined by the strength of the springs, *it is the springs that determine the inertia*.

For example, in classical physics the amount of energy you put in to move something is called “work”. When a constant force is applied, we have

$$\text{work} = \text{force} \times \text{displacement} = \text{mass} \times \text{acceleration} \times \text{displacement}$$

So, when we divide the work by the mass of the object, we get

$$\text{work/mass} = \text{acceleration} \times \text{displacement}.$$

Hence, the acceleration is determined by the ratio between the energy put in (the work) and the mass of the object. Consequently, A massless wave (only vacuum elasticity) requires no energy to reach light speed. An electron (higher elasticity) can be accelerated to (almost) light speed using a fair amount of energy. But something as heavy as a spaceship cannot be accelerated to light speed by any means available to us today. It requires too much energy.

We can also clarify this in a different kind of picture (see Fig. 6.5). Here, the total energy is on the vertical scale. The momentum is on the horizontal scale. In the picture you see that when the momentum is 0, the energy is determined entirely by the mass of the particle/wave. For massless particles it is 0 (as for light), while for massive particles it is $M_0 C^2$. When the momentum goes up, the massless particles follow the straight line $E = PC$. The massive particle follows the line that is dominated by the springs at low momentum, i.e., $E = MC^2$, and the line that is dominated by the rope at high momentum, i.e., $E = PC$. And so, we see that this line is not straight, but bent like a shell. Now imagine that the momentum can be in three dimensions (the three directions we can move in) and you will see that the line of the massive particle becomes a sort of cup or shell. This is called the *mass shell*. As we will see later a real particle is said to be “on-shell”, which means that it follows this energy line properly. A virtual particle is said to be “off-shell” which means that it deviates from this line. But that’s for later.

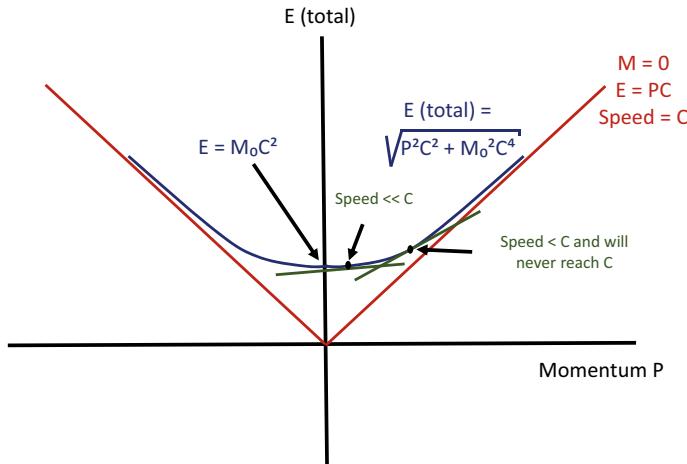


Fig. 6.5 The mass shell. The (group!) velocity of a quantum wave is determined by the way the total energy (frequency) changes with the momentum (wavelength), hence by the slope of the graph. For massless particles the slope is constant: the total energy equals the constant C times the momentum. For massive particles, the total energy includes the energy that goes into the mass (the oscillation of the springs). Hence, the way energy (frequency) depends on the momentum (wavelength) is not constant

When energy is fed to a massive particle, it increases both its velocity and its momentum. At low momentum values, we should find the classical formulas for energy and momentum. When the velocity gets close to the speed of light, the velocity cannot go up much further. However, the momentum still increases if we keep adding energy to the particle. This means that the momentum cannot go up in proportion to the velocity (as in the classical formula $P = MV$).

We can also see this in the following way: the energy put in leads to an increase in frequency. Initially, this leads to an increase in velocity and a decrease in wavelength according to $F = V / \lambda$. With increasing velocity, the momentum goes up, which means (since $P \sim 1/\lambda$) that the wavelength gets shorter. So, at still lower velocities, the increase in frequency gets spread over an increase in velocity and an equal decrease in wavelength. When the velocity increases until it cannot increase much further (being close to C) the further increase in frequency can only be translated into an extra decrease in wavelength (see Fig. 6.6).

Figure 6.6 shows the dispersion relation of various waves. Let's first look at the two sketched situations for low speed on the left of Fig. 6.6 and high speed on the right:

a. The left-hand side of Fig. 6.6 (low speed)

When the kinetic energy is much lower than the mass energy, the wave behaviour is dominated by the mass energy. Hence, we should get the same behaviour for the group velocity as dictated by the classical formulas for particles, i.e., we should get $E = \frac{1}{2} PV$ and $P = MV$. Let's now double the energy. This would mean that the

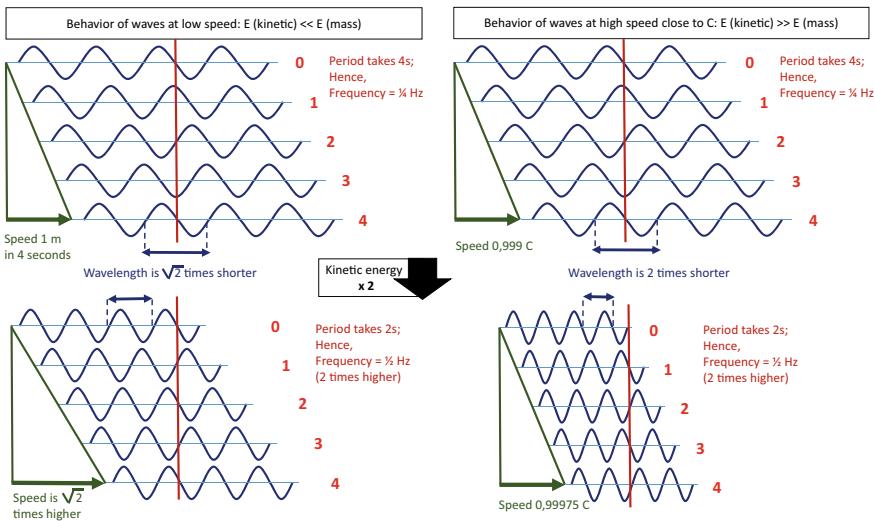


Fig. 6.6 The dispersion relation of massive waves at low kinetic energy (left) and at high kinetic energy (right)

frequency of the wave gets doubled as well. When we spread the increase in energy equally over the particle velocity V and the momentum P , we find that both classical formulas are satisfied:

$$2E = 2 \times 1/2PV = 1/2(\sqrt{2}P)(\sqrt{2}V)$$

$$\sqrt{2}P = M\sqrt{2}V$$

So, when the energy (frequency) gets doubled, the momentum gets increased by a factor $\sqrt{2}$ and the velocity gets increased by the same factor. Consequently, since momentum $\sim 1/\lambda$, the wavelength gets reduced by a factor $\sqrt{2}$, which is in line with $F = v/\lambda$.

b. The right-hand side of Fig. 6.6 (high speed)

When the kinetic energy is much higher than the mass energy, the wave behaviour is dominated by the kinetic energy. In this case too, the frequency gets doubled when the energy gets doubled (as $E \sim$ frequency). However, the velocity hardly changes, so doubling the frequency leads to a halving of the wavelength, or

$$2E \sim 2F \sim 1/2\lambda \sim 2P$$

So, doubling the energy means doubling the momentum. Since the velocity is almost C , we see the relation $E = PC$ appear. In this case, group velocity = phase velocity = C . We will investigate the difference between the left- and right-hand

sides of the picture in more detail in Sect. 6.4. Then we will discover how this effect is related to length contraction and time dilation.

Conclusion

When the mass energy dominates the behaviour of the wave (low velocity), a doubling of the kinetic energy gets equally spread over an increase in momentum and an increase in velocity. In this case, the classical behaviour prevails. However, when the kinetic energy dominates the mass energy, the relativistic behaviour appears (velocity close to C), like the behaviour of massless waves in vacuum. The fun is that all this behaviour can be explained by considering waves ($F = v/\lambda$) and a simple picture of springs attached to a wave.

6.2.5 Consequence 4: Other Potential Differences Can Create “Mass”

It is not only the Higgs field that gives mass to particles. For example, an atom has a lower mass than the sum of the masses of protons and neutrons. You must put energy in to break the atom apart. So, when we bring the protons and neutrons to a certain distance, they possess potential energy. What is this energy? Basically, they feel a potential difference. The potential difference causes the particle-wave to have a higher mass energy, just as happens with the springs.

However, this potential difference is not the same as the Higgs field. The Higgs field has the same strength everywhere, so it does not provide a potential difference *in space*. Consequently, no particle will experience a force from the Higgs field as forces only result from potential differences *in space* (as we will see in Sect. 9.4). The Higgs field can only be noticed if you try to excite a field that connects to Higgs. In that case, the springs from the Higgs field cause you to put a lot of energy in before you can excite the field and create a wave (a particle).

The potential of the strong force (as applicable in the example of the atom) does show a potential difference *in space*. It is that potential difference that causes a force to be felt between the protons and neutrons of the nucleus. So, the springs are there, but they are just arranged differently (see Fig. 6.7). And they too change the behaviour of the wave by contributing to its inertia. In the picture the force is represented by springs. It is true that the potential difference is higher when the distance to the source of the force is larger, but the analogy breaks down when one compares the force of, e.g., the electromagnetic field with that of a spring. The spring will exert a greater force when stretched, while the electromagnetic field will exert a smaller force at greater distance.

Take for instance the He^4 nucleus. When we compare its mass to the total mass of 2 individual protons and 2 individual neutrons that constitute it, we get an interesting result (see Table 6.1).

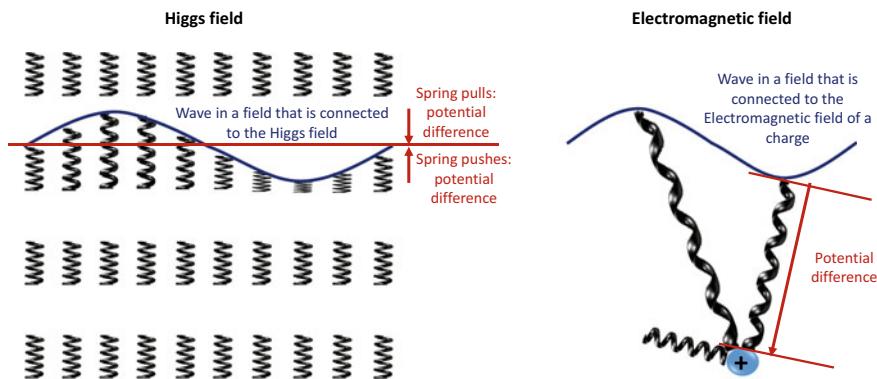


Fig. 6.7 Higgs and other fields are different and each produces a potential difference in its own way. A wave in the field that experiences these potential differences becomes massive with higher elasticity, higher inertia, and lower velocity

Table 6.1 The mass of a nucleus

Proton mass	$1.672621898(21) \times 10^{-27} \text{ kg}$	$\times 2 = 3.345243796 \times 10^{-27} \text{ kg}$
Neutron mass	$1.674927471(21) \times 10^{-27} \text{ kg}$	$\times 2 = 3.349854942 \times 10^{-27} \text{ kg}$
Total of 2 neutrons + 2 protons		$6.695098738 \times 10^{-27} \text{ kg}$
Helium nucleus mass		$6.644657230(82) \times 10^{-27} \text{ kg}$
Difference		$0.050441508 \times 10^{-27} \text{ kg}$

So, we see that the mass of the helium nucleus is some $0.05 \times 10^{-27} \text{ kg}$ lower than the mass of the individual protons and neutrons. If we could take the individual protons and neutrons to a certain distance from one another, they would gain a mass of $0.05 \times 10^{-27} \text{ kg}$. This mass is created as “spring energy” or potential energy (from the potential difference) as experienced by the proton and neutron waves.

So, springs, Higgs and potential difference are the cause of mass. Mass is nothing else than the change in behaviour of the wave because “springs” are attached to it, or better, because the wave is connected to a field that creates some kind of potential difference. But what is a potential difference? How does a field get attached to another field and how does it make the field feel like it is attached to springs? We will try to answer all these sorts of questions in the following chapters. We will be looking at how fields interact and the concept of virtual “particles” (waves, of course...). We will also be looking at how the wave is influenced during its path through space using the path integral, and we will be looking at how a force is created. By then you will

have some idea of what we mean by a potential difference and fields “connecting” to other fields.

6.2.6 Consequence 5: Mass Can Be Changed Into Energy

When the protons and neutrons from the previous example are released, they gain velocity as they get together again in the He⁴ nucleus and the potential difference is reduced. So potential energy is exchanged for kinetic energy. When a neutron wave experiences a lower potential energy because it gets closer to the other neutron and protons, it starts to behave as though it feels less spring strength. Consequently, it behaves as if it has a little less mass. At the same time, the total energy (frequency) remains the same. Consequently, the energy has to go “somewhere”. What happens is that the wave in the rope starts to get a higher velocity, when the potential energy decreases. In terms of frequency, some “mass frequency” is exchanged for “momentum frequency”. You can imagine from our discussion so far how a lower spring strength decreases the elasticity. A decreased elasticity leads to a higher velocity (see Fig. 6.2).

One may argue that the energy that is in the momentum of the particle has a mass equivalent. Hence, the “mass” energy that changes into momentum energy does not really change the total mass. The total energy remains the same, hence the total mass remains the same. However, the momentum energy can be transferred to other particles! When that happens a lower mass energy remains in the original particle. So, when the protons and neutrons from our example start to gain momentum when they get together, that momentum may be transferred to other particles. When that has happened, we have a He⁴ nucleus that has a lower mass than the original neutrons and protons.

Particles can be annihilated as well. For instance, when a particle meets an anti-particle. Then the total mass of both is destroyed and changed into momentum energy, e.g., by producing two photons. We will see later how this works. In short, the spring frequency is completely stopped and momentum frequency of a different field is produced instead.

6.2.7 Example: Photons in a Plasma

We have worked out what the consequences are when waves enter a different medium. An interesting example is the behaviour of photons in a plasma. A plasma is a gas in which the electrons are separated from the atoms. Consequently, they are free to interact continuously with photons at any time and with any energy. Hence, a plasma acts with regard to the photons as a continuous and homogeneous interacting field. The interaction the photons have with the plasma acts like springs on the rope. The

fun part is that when this is worked out [Ref. 54, handout 14], the massless photons have exactly the same behaviour as a massive particle in vacuum!

Inside the plasma, photons below a certain energy cannot exist. You can compare the situation below that energy threshold with moving the rope up and down too slowly in Fig. 6.2. In that case, the potential pulls the rope down faster than you do. The consequence is that no wave gets started. Only by moving faster than the potential could induce can a wave get started. Hence, there is a minimum energy threshold that plays the role of the mass. Above that threshold, photon waves can only move with a group velocity. That velocity depends on the energy of the photons relative to the threshold energy! So, when the energy of the photons is only very slightly above the threshold energy, their group velocity is very low. When that energy increases, so does the group velocity. When the group velocity gets close to the speed of light, the relation between the photon energy and the threshold energy obeys exactly the same formula as the relativistic formula for massive particles. That means that the velocity can come close to the speed of light, but when it does, it takes an increasing amount of energy to increase the velocity of the photons any further. From this we can conclude that the threshold energy does indeed play the role of mass. This comparison tells us that, when we view massive particles as waves with springs attached, they will indeed exhibit all the behaviour we are used to from massive particles.

6.2.8 Conclusion and Summary

We have developed an idea of how we can see a particle as a wave packet, a group of waves. Such a group of waves can exhibit all the behaviour of a massive particle when the waves interact with a continuous field (a Higgs field), by analogy with springs on a rope. This picture explains the relativistic behaviour of “massive particles” close to the speed of light.

In that behaviour we have identified two kinds of energy:

Momentum energy

Momentum energy is stored in the elasticity of the medium. When a wave is created in a medium such as a rope or the vacuum, it holds a certain amount of energy. This is called the kinetic or momentum energy of the wave. The frequency of the wave is equal to its energy, the wavelength is inversely proportional to its momentum, and frequency and wavelength are related to each other via the velocity of the wave.

Mass energy

Mass energy is related to springs that can be attached to the wave. These springs can be actual springs attached to a rope or a potential pulling a wave in the vacuum. The springs can store energy as well, and this is called mass energy. This can be a Higgs potential giving rest mass to the wave, but it can also be a “regular” potential created by another field giving mass to the wave that we interpret as potential energy. The

mass energy changes the elasticity the wave experiences, and hence also the relation between frequency, wavelength, and velocity.

These changes influence the way the velocity of the wave can be changed by a force. When the springs are stronger, more energy is stored in them and it becomes more difficult to change the velocity of the wave packet: the energy stored in the springs creates the notion of inertia.

Mass energy can be exchanged for momentum energy. The energy stored in the Higgs springs attached to the wave (the wave's rest mass) can only be exchanged for momentum energy when the wave is annihilated. This is a consequence of the fact that the Higgs springs are there as soon as the wave is created, and they are always equally strong. A regular potential changes its potential difference depending on the distance to the source of the potential. So, a regular potential can be changed continuously into momentum energy when the distance to the source is changed (the concept of force).

So, mass/inertia, energy, and velocity are all consequences of the behaviour of waves in the elastic vacuum and of the springs (potential caused by another field) that get connected to the wave and change the experienced elasticity.

This is quite spectacular! What we have done here is to explain the origin of mass and inertia from a wave perspective and to explain why the velocity of a massive object behaves in the way dictated by special relativity.

6.3 The Elasticity of the Minkowski Metric

What is a metric? A metric describes how space–time behaves. For instance, a flat piece of paper can be described by perpendicular straight lines. But a balloon can best be described by bent lines (such as longitudinal coordinates, as used on earth). Both these shapes can be characterized by the definition of “distance”. Any distance on the flat paper is described by Pythagoras: $\text{distance}^2 = x^2 + y^2$ if we call the two possible perpendicular axes that describe the paper x and y. On the sphere, we get a spherical formula to describe a distance. You cannot go straight from A to B when you are on earth, you have to follow the curved surface. When you fly to the other side of the earth, you cannot go straight through the core of the planet, you have to fly around. So, distance is defined differently on a sphere than on a flat piece of paper.

Now what about space–time? In our daily lives, we would define a distance in space in the same manner as on the piece of paper, using Pythagoras: $\text{distance}^2 = x^2 + y^2 + z^2$. But then we have not included time! So how to deal with time? Including time, we would define a distance as the “separation” between event A (at position X_a, Y_a, Z_a and time T_a) and event B (at position X_b, Y_b, Z_b and time T_b). But what is the distance between T_a and T_b ? If we want to measure this as a distance, we have to multiply the time difference by a velocity: $v(T_b - T_a) = \text{distance}$ between the two events in time only. But what velocity to use here? It is useful to choose the maximum velocity available, which has the same value for everyone. This gives us the minimum distance in time: it is the fastest way to get from A to B. After all, when

we measure distance on a piece of paper, we don't go around in loops either, we take the shortest distance between two points to define "distance".

So, what about this maximum velocity? The vacuum provides us with a maximum velocity as a consequence of its elasticity. So, distance in time in the vacuum should be measured as $C(T_b - T_a)$ or in short CT . The consequences of this are far-reaching. The German mathematician Hermann Minkowski was the first to define this metric. It describes distance as follows:

$$\text{Distance}^2 = (CT)^2 - (X^2 + Y^2 + Z^2) = (CT)^2 - X^2 - Y^2 - Z^2$$

This way of calculating the distance in four-dimensional space-time is called the Minkowski metric, and it is the basis for all the formulas and behaviour encountered in special relativity. We call the space-time that is built up using this metric Minkowski space.

This metric has some interesting features:

a. The distance is 0 when one moves at the speed of light

Just do the calculation, moving in the X direction (so $Y = 0$ and $Z = 0$) and travelling at 300,000 km in 1 s: so, $x = 300,000$ km and $T = 1$ s:

$$\text{Distance}^2 = (300,000 \text{ km/s} \times 1\text{s})^2 - (300,000 \text{ km})^2 = 0$$

This means that when you fly with the speed of light any distance is 0. Put differently, you would not age when you move from A to B; you would immediately get there! Even though other observers see you fly by at a velocity C. Suppose you fly at the speed of light to the nearest star some 4 light-years away. An outside observer would say that you took 4 years to get there. But you yourself would start off and in literary *no time* you would be there. It is rather as though the speed of light has taken the place of a kind of "infinite velocity". As we will see later, this can only work if the length (distance in space alone) gets contracted and time gets dilated. All this as a consequence of the elasticity of the vacuum!

b. The distance between A and B can be time-like or space-like

One can only travel from A to B when the distance in space is shorter than it would take light to get there. For instance, when you have to move from where you are sitting to the window and you have an hour to do it, you will be able to do it. This is called a time-like distance. However, when you are asked to go to Mars in one second (from the perspective of an outside observer), you cannot do it, and you could not even if you could move at the speed of light. This is a space-like distance. So, a time-like distance describes a separation that you can cross, while a space-like distance is one that you cannot cross.

The surface separating distances you can and cannot cross is called the light cone (see Fig. 6.8). The light cone links all points that can just be reached by light. The dark blue highlighted area is the space-like area that one cannot reach.

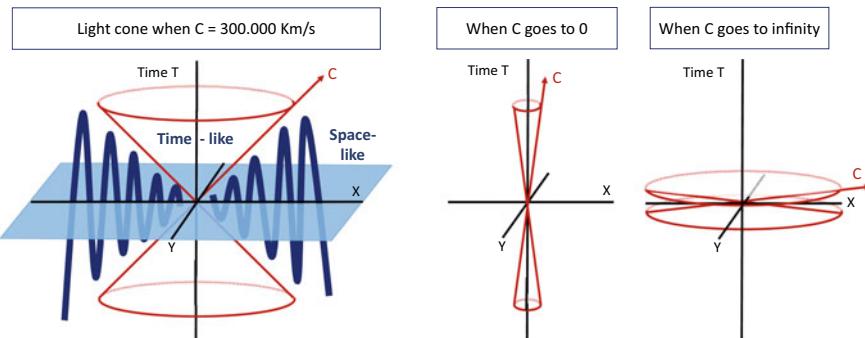


Fig. 6.8 The light cone

c. If C were 0, there would effectively be no space and time

When $C = 0$, $V = 0$. So, you cannot get anywhere. All other places in the universe (except those on your own timeline) cannot be reached. So, we would not know of any other particle even if it was sitting right next to us. The light cone looks pinched (see Fig. 6.8). The time-like region is reduced to nothing.

The elasticity of the vacuum would be very high. In fact, exciting one point in space-time, would not lead to any excitation of the next point in space-time (the vacuum is too elastic). Consequently, no wave could exist. This implies that there could be no knowledge of other points in space-time and effectively one could say that distance and other points in space-time did not exist.

If we could consider C to be, not 0, but close to 0, let's say 1 m/s, a wave would be possible. However, the vacuum would be so elastic, that a spring with the slightest spring strength would cause the waves to behave differently (see Fig. 6.2). The mass energy associated with that would be very low. If we consider a mass energy from our everyday lives in our own universe, it would completely change the behaviour of waves and it would be very difficult to get any speed at all. Just imagine pulling up a very elastic rope with heavy springs attached. You would pull it up, but the rope would not have enough power to lift the next spring. The rope would simply stretch out. The only way would be to pull the rope up so fast that it did not have time to stretch out and might thus lift the next spring. So, in order to get a wave going, one would need to put in enormous amounts of energy to overcome the springs.

If one managed to put in such energy and get a wave going with some (low) speed, one would quickly get relativistic effects. For instance, when you got on the train to work, your watch would be behind compared to those that stayed home. Communicating with the ones that stayed home would be difficult, since it would take any signal a long time to get there and back.

d. If C were infinite, space and time would be endless

When C is infinite, V can have any value (see Fig. 6.8). An infinite C implies that the elasticity of the vacuum is 0. It is totally inelastic, so any trigger of a massless

field in such a vacuum would spread immediately at an infinite velocity. To change the behaviour of such an inelastic vacuum (i.e., making it more elastic) one needs infinitely strong springs (see Fig. 6.2). This is difficult to imagine, so let's take a very large value of C (but not infinite). One would need very strong springs to make such a vacuum a bit more elastic, but not infinite. Suppose one has such springs, the mass energy would be extremely large and it would be very hard for kinetic energy to dominate it. Consequently, the behaviour of massive waves would be like massive waves in our world at low velocity. Essentially, in such a vacuum you could have any velocity and an increase in kinetic energy would lead to an increase in velocity. The behaviour of such massive waves would almost never become like $E = PC$ (see Fig. 6.6).

None of the relativistic effects as we know them in our world would exist for very massive waves. There would be no time dilation and no length contraction (see Sect. 6.4). It would be possible to have a very high velocity without aging more than those that stay behind. We would in principle be able to get to the nearest star while everyone else, including those that stay home, would be able to experience the result and see a live recording of the event.

So, we can conclude that it is the elasticity of the vacuum that determines how our waves behave and how difficult it is to get somewhere else. It is the elasticity of the vacuum that determines what relativistic effects we experience. This is not unexpected. After all, if we are all made of waves in the medium that we call the vacuum, the way we experience space–time must be determined by how these waves behave in the vacuum.

6.4 Length Contraction and Time Dilation of Waves

In Sect. 6.2 we saw that, when the velocity of a massive wave gets near C , its velocity cannot increase much further. When we keep increasing the energy (the frequency) of the wave, the velocity cannot go up, so the wavelength has to get shorter than expected (see Fig. 6.6), in accordance with $F = V/\lambda$.

In Fig. 6.6 we saw that this leads to an extra decrease in wavelength compared to the low velocity scenario. What could the extra decrease in wavelength originate from? After all, for the waves there is only one relation between frequency, velocity, and wavelength (the formula mentioned above). So, an “extra” decrease in wavelength cannot be explained from the wave behaviour. There is only one way to solve this paradox. It must be space itself that contracts.

How can we see that? The extra decrease in wavelength we need is shown in Fig. 6.6. We can call this a mismatch in the formula that relates frequency to wavelength for massive waves with a velocity close to C . But space and time must be treated on an equal footing. Consequently, the mismatch between frequency increase and wavelength decrease (when V stays approximately the same) cannot be completely compensated by the contraction of space alone. So, what must happen on the time side to compensate the mismatch? Time is related to frequency, so compensation on

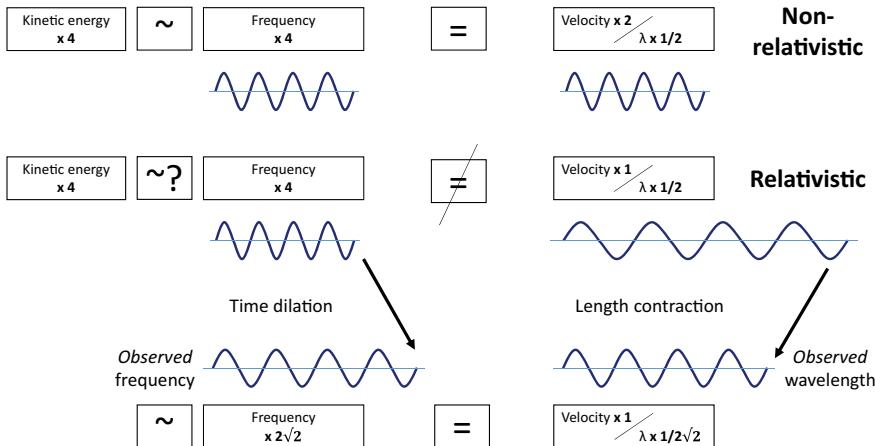


Fig. 6.9 The mismatch between wavelength and frequency in the relativistic case is resolved by time dilation and length contraction. On the left-hand side of the equation, you see that the frequency gets contracted (multiplied by a smaller number) and the period gets dilated. On the right-hand side of the equation, the wavelength is divided by a larger number and so it gets contracted, so the formula gives the correct result. Please note that the frequency does not really get multiplied by a smaller number, but it is time that dilates so that any outside observer would observe a slower clock and consequently observe a lower frequency

the time side should result in contracting the frequency as much as we contract the wavelength. When we do that, the mismatch is completely compensated and on an equal footing (see Fig. 6.9).

Now you might be wondering what frequency contraction means. The frequency of a wave is defined as 1/period. The period (T) is the time that passed between the start of a wave and the completion of one wavelength. So, $F = 1/T$. Consequently, when the frequency gets contracted, it must mean that the period gets dilated. This is called time dilation. Everything will go more slowly when the velocity approaches C . For instance, if $T = 1 \text{ h}$ and we look at the clock to see when 1 h has passed, we find at high velocity that T takes a much longer time, as observed by someone who sees us rushing by. Close to C , our clock goes extremely slowly, our heartbeat looks as though we have died, and we simply won't age.

So, we conclude that for massive waves the ratio between kinetic energy and mass energy determines the velocity of the waves as well as the way space and time behave. At low kinetic energy the mass energy dominates. The formula $F = V/\lambda$ works and there is no mismatch. At high kinetic energy, there is a mismatch: the formula $F = V/\lambda$ only works when both the frequency F and the wavelength λ are contracted on an equal footing so V does not have to change much when the energy increases.

Ultimately, when the mass energy is negligible compared to the kinetic energy, the behaviour of waves should look like that of massless waves. The relative impact of the springs becomes low at such high kinetic energies. This means that the Minkowski distance must come close to 0, as is the case for massless waves such as light. There

is no length and the time passed between two events on the world line of the massless wave is 0. So, its clock would seem to go infinitely slowly. As we saw before, the formula for the wave will approach $F = C/\lambda$.

For massive waves at low velocities the distance is not 0. Hence, giving mass to a wave not only changes its properties, *but also creates the notion of distance in space and time*. Moreover, velocity does not in fact exist for massless waves. If our world contained only massless waves, everything would happen immediately and no-one would be around to observe “the speed of light”. So, velocity is also a concept that goes together with a wave becoming massive.

Length contraction and time dilation are characteristics of the vacuum that relate the massive world with distance and time to the massless world where distance is 0 and time does not go by. It is an effect that comes with changing the elasticity of the vacuum using a potential such as Higgs.

Summarizing:

1. Springs on a wave change the elasticity.
2. At low momentum the mass aspect of the springs dominates the wave behavior. The relation between wavelength and frequency follows the dispersion relation, meaning that an increase in energy gets spread over an increase in speed and an equal decrease in wavelength.
3. At high momentum (high speed), the relation between wavelength and frequency increasingly starts to look like that of a massless wave, with a constant speed close to c . An increase in energy can no longer lead to a substantial increase in speed.
4. As seen by an outside observer, this makes speeding waves adjust to the massless wave behavior by shortening their wavelength and prolonging time intervals.
5. Hence, relativistic effects such as time dilation and length contraction are nothing else than our waves shifting from predominantly springlike behavior towards primarily wave behavior.
6. Since everything is made of waves, including clocks and measuring rods, everything shows length contraction and time dilation as a consequence of our waves adjusting to higher speeds.
7. The result is that at low speed we have a different perception of distance in space-time compared to our perception at high speed.
8. Consequently, the notion of distance in space-time is created by the changed elasticity caused by the springs. A large distance is experienced at low speeds, but a short distance at high speeds.
9. Since we can only compare ourselves to other waves (with which we interact), these effects are always relative between different frames of reference, i.e., between massive waves moving at different speeds.
10. An observer can see a speeding massive wave have a speed close to c and a light wave at c . Yet, the speeding massive wave will see the light wave move at c as well. The observer can explain this difference in perception from the contracted measuring rod and dilated clock of the speeding massive wave.

11. When the observer increases his speed to that of the massive wave, he will have the same contracted rod and dilated clock. Consequently, he will still see the light wave move at c and he will see the clock of the massive wave go round in the same time interval as his.
12. All this is the direct consequence of the interplay between waves and springs!

6.5 About Higgs

We have already mentioned the Higgs field on several occasions, and we will come back to it on several more, but let's do a quick preview.

The Higgs mechanism is often explained, e.g., in popularizing blogs, in the following popular way:

- One example is the “molasses” analogy. In this analogy, the Higgs field is compared to molasses that fills the space. Particles that interact with Higgs are envisioned as being slowed down by the molasses so their velocity is no longer C and they get an inertia. That is, accelerating such a particle in molasses takes energy. The analogy immediately breaks down when one tries to envision how such a particle keeps its velocity when no force is applied. Of course, the molasses would slow it down until it stops. So, this does not give me a great feeling of understanding the matter.
- Another example is the “celebrity” analogy. In this analogy, the particle is compared to a celebrity in a room full of fans. Here too, the particle is slowed down by fans trying to get an autograph. Accelerating such a celebrity particle is difficult, hence its inertia. But this analogy does not explain how the particle can keep its velocity either.

When we take the view as presented in this book, where the particle is a wave, and the Higgs field is like springs connected to the wave, we get a better analogy. It is clear why the velocity is lower and how it gets inertia. Moreover, the wave will keep on going when no force is applied. However, our view does help to understand what the other analogies actually mean by slowing the particle down. The essence of all the analogies is that the all-pervasive Higgs field connects to the particle wave, which changes its properties, slowing it down to below C and making it harder to accelerate it.

Our view is also closer to the math, as it is described there as a field that connects to the particle field that gains mass from the Higgs potential. We will come back to the Higgs field in Chaps. 16 and 17 when we discuss broken symmetry. Then we will say a bit more about the origin of the Higgs field.

So far, we have said that the mass of a particle depends on the strength of the springs, i.e., the strength of the connection between the particle wave and the Higgs field. However, this does not explain why we cannot create a heavier electron, for instance. We just need to put in more of an electron wave and we are done! So why do electrons (and all other particles) come only in one size? Using our understanding

of energy and mass in wave terms we can now look at how the fields in the vacuum can get quantized into one size. We will address this in the next chapter.

Chapter 7

Quantization of Fields



We have seen so far that particles can be represented by waves. The process by which particles are described as waves is called first quantization. Why is this called quantization? A “quantum” can be defined as a minimal amount of a physical entity involved in an interaction. In physics, the idea that physical entities may be quantized (discrete rather than continuous) is called the “quantization hypothesis”.

7.1 First Quantization

Quantization is directly related to the wave character of particles. For instance, if you try to fit a particle in a box and a particle is actually a wave, a condition arises. The wave itself must be continuous. So, if we look at the edge of the box the wave cannot have a nonzero value, while outside the box it must be 0. So, the condition we get is that the wave must be 0 at the edge of the box (see Fig. 7.1).

If the wave must be 0 at the edges of the box, it can have only particular wavelengths: the wavelengths that fit in the box. So, the wavelength must be “quantized”: only discrete values of the wavelength are allowed. Since the wavelength is directly related to the frequency and hence the energy, those must also be quantized. The same goes for the momentum. So, we see that a particle wave in a box can only have discrete energy levels. In nature we see this in atoms. Electrons in atoms can only have discrete energy levels. We will go more deeply into this in Sect. 13.1 when we discuss spin.

So, representing particles by waves leads to quantization of their energy and momentum in particular situations.

The idea of a box with infinite potential is of course rather academic. Just a little less academic is a box with a finite potential. This allows an interesting possibility: the uncertainty relations allow a wave to have some extra energy for a short while. This would mean that the wave does not have to have 0 amplitude at the wall, but could have some amplitude there and extend beyond the wall (see Fig. 7.2).

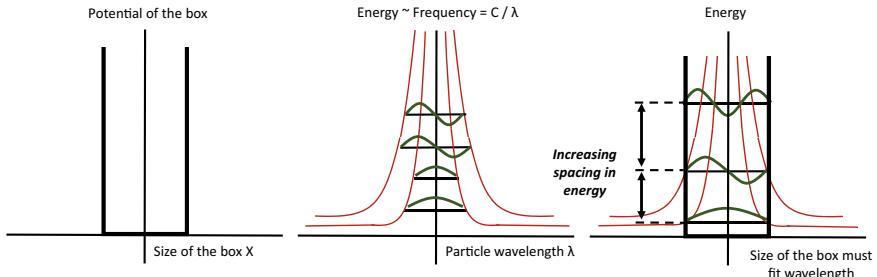


Fig. 7.1 On the left is a box with impenetrable walls. We could describe an impenetrable wall as a wall with infinitely high potential at the edge. Nothing can cross an infinite potential, so the wave amplitude outside this wall must be 0. The middle graph (red) shows how the energy of waves (green) depends on their wavelength. The third graph shows the first two graphs laid on top of each other. If we fit half a wavelength in the box, we must look at where the “half wavelength” red graph crosses the box walls. Then we find the energy that goes with that situation. If we fit one wavelength in the box in the same way, we find a higher energy level. It turns out that as we go up each energy level exhibits a greater difference in energy than between the previous levels

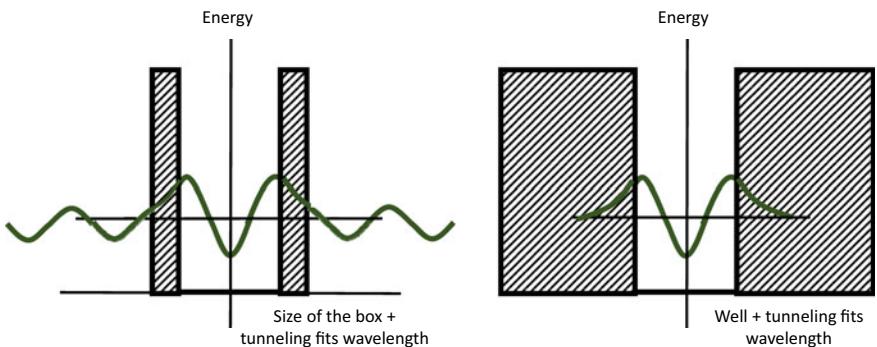


Fig. 7.2 Tunnelling through the potential wall of a box (left) or a well (right)

The picture shows two possibilities. Option 1 is a potential wall. Beyond the wall the potential is 0 again. In that case, the wave can “borrow” extra energy for a short while to overcome the wall and continue as a wave beyond the wall. There is a small but non-zero probability for this process to happen. The field is lower beyond the wall, indicating that the probability of finding the particle there (i.e., of interacting with the field) is lower.

Option 2 is that the box is essentially a potential well. The potential remains high beyond the wall. The wave can borrow extra energy but only for a short time. So, the wave cannot sustain itself outside the well. It can only penetrate a short distance. So, the wave quickly declines to 0 beyond the wall.

All this can be seen from the perspective of the previous section, where a potential is thought of as mass energy. Within the box the wave simply does not have enough (kinetic) energy to trade in for potential energy (mass energy). But since the box

has a limited size, the particle wave must be built from a number of waves, based on the uncertainty relations. Some of these waves will have the energy to swap for mass energy, representing the possibility of crossing the wall. This process is called tunnelling.

7.2 Second Quantization

Quantum field theory proposes a “second quantization”. In this case it is not particles that are waves with a quantized energy and momentum (“first quantization”). Now we are looking at waves in a field and concluding that they can be particles, so it is rather the other way around.

The point here is that particles seem to show up in nature with a fixed rest mass. The rest mass of an electron (the mass at 0 velocity) cannot be changed. This looks a lot like quantization. You can have one electron or two, but nothing in between. We cannot take 1.5 times the energy required to make an electron and decide to make 1.5 electrons. It is almost like the waves in a box, leading to discrete energy levels.

So, let’s see how this works. We said that a particle is really a wave in a field. We have seen in the previous section that the vacuum is a medium that can carry waves (like a rope). We also saw that when such a wave is massive, we can consider it as a rope with springs attached to a wall. So, let’s take a closer look at those springs.

What is a spring? When it is stretched and released, it will jump back. If there were no energy loss in the process, the spring would go back to its original position, gaining velocity and overshooting its original position. Until it gets squeezed, after which it starts to stretch again. Without energy loss, this would continue forever. A little like the weight of a standard clock on the wall. This type of movement is a harmonic oscillator, as discussed before. It oscillates at a particular frequency.

When we do this to a spring and it starts to oscillate, we essentially store energy in it. It oscillates at a frequency consistent with that energy. It will do that until we stop it by touching it. If we do so, we feel the bump of the spring against our hand, which means it transfers its energy back to our hand.

As we stretch the spring further, it becomes harder and harder. The more we stretch it, the more energy it takes. In fact, this is not linear: when we stretch it twice as far, it requires $4\times$ as much energy. Moreover, the spring’s strength plays a role. When the spring is twice as strong, it takes twice as much energy to stretch it over the same distance. This is Hooke’s law (see Fig. 7.3).

When we want to excite a field with springs attached, we have seen that we first must put enough energy into the spring. We called that the mass energy. Now let’s see what that really means for a field. We saw that the Higgs field plays the role of the springs. Suppose the field gets stretched and it feels the Higgs field pulling back. When it is stretched to a certain field amplitude, we know that there is 0 probability of finding the field stretched beyond that. So, when we use a wave to describe the oscillation process, it must be 0 at the greatest stretch.

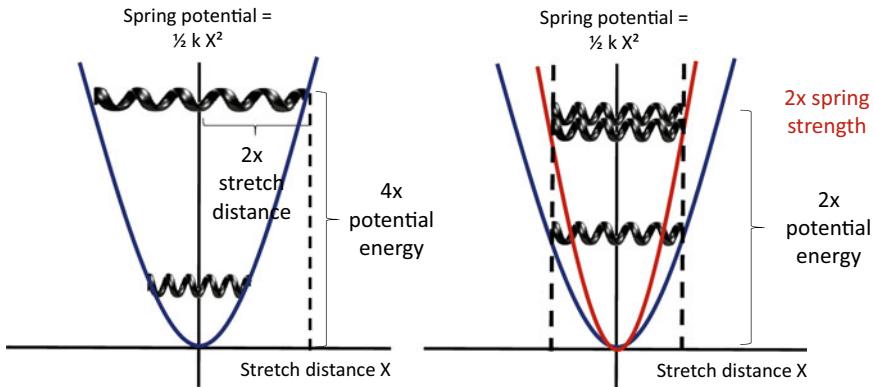


Fig. 7.3 Hooke's law: the potential energy in a spring increases with the square of the stretch distance: stretching it twice as far requires four times the energy. Moreover, an important factor is the strength of the spring itself. On the right we see what happens when the spring is twice as strong: the potential is narrowed. This means that it takes twice as much energy to stretch it over the same distance

You can see where this is going: when a wave must be 0 at both maximum amplitudes (see Fig. 7.4), there can be only a discrete number of wavelengths allowed. Only those wavelengths that fit this amplitude. Now we are “second quantizing” the field.

Although we can compare this to what happened with the waves in a box, the spring is not a box. It does not have an infinite potential at a particular distance. The potential of the spring is determined by Hooke's law. The further the spring gets stretched, the higher the potential. It is not trivial to see what happens to the wavelengths that are allowed. So, let's draw a picture (see Fig. 7.5).

The figure shows first (on the left) the spring's potential. The horizontal axis represents the field amplitude (the distance the spring is stretched). On the vertical axis the potential energy stored in the spring (equal to the energy put in to stretch it). The middle graph in the picture shows how waves are related to energy. The higher the energy of a wave, the shorter its wavelength ($E \sim F = v/\lambda$). The nature of this

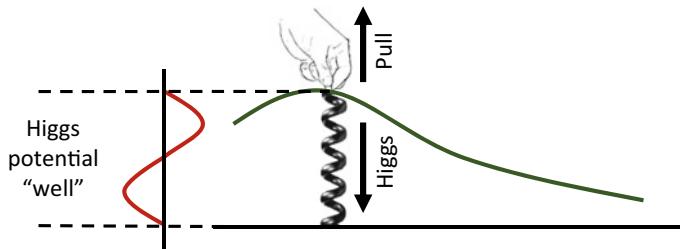


Fig. 7.4 Attaching a wavelength to a stretched spring

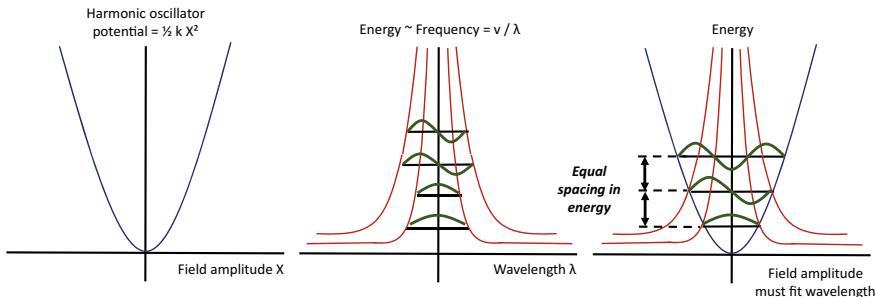


Fig. 7.5 The energy required to excite the field is related to a frequency. That frequency is related to a wavelength. The relation is such that a higher energy leads to a shorter wavelength ($F = v/\lambda$). This wavelength must fit the field amplitude. The first graph shows the field potential (blue) as a function of the field amplitude. The second graph shows the energy of waves (red) as a function of their wavelength. The third graph shows the two graphs laid on top of each other. If we fit half a wavelength into the field amplitude, we must look at where the “half wavelength” red graph crosses the blue graph of the field amplitude. Then we find the energy that goes with that situation. If we fit one wavelength into the field amplitude in the same way, we find a higher energy level. It turns out that each energy level up is equally spaced

equation is that when the wavelength gets really short (close to 0) at constant velocity, the frequency must become really large (tending to infinity as λ goes to 0). The inner red graph shows how half a wavelength shortens when the energy (frequency) goes up. The outer red graph shows the same for one wavelength.

So now we have to fit such a wavelength into the amplitude of the springs attached to the field. This means we have to put both the left and middle graphs on top of each other so that we can fit the wavelength into the field amplitude. Hence, we get the graph on the right. Where a red graph crosses the blue graph, the wavelength (red graph) is equal to the field amplitude (blue graph).

This gets interesting when we look at the energy levels at which this takes place. These energy levels turn out to be exactly equal in spacing. So, the energy difference between fitting half a wavelength and fitting one wavelength into the field amplitude is exactly the same as the energy difference of fitting one wavelength and fitting 1.5 wavelengths.

So, it turns out that we can excite a field with springs attached and store energy in it only by discrete amounts. The minimum level to which we can excite the field is the level that corresponds to fitting in half a wavelength. When we try to put more energy in, it cannot lead to a higher field amplitude until we put in so much energy that the field gets excited by fitting one wavelength in.

Exciting the field by half a wavelength produces what is called the ground state of the field. In fact, the field cannot have less energy! So, any field with springs attached must have a minimum energy or it cannot exist. When we put energy into a field, the first level we can excite it to is the level that fits one wavelength. That level is called one particle. We either put energy in, equal to the mass of the particle, or the field

does not get excited to that level. Period. You cannot excite the field halfway or in any continuous way.

What does it mean to have two particles in the field? It basically means that the field is excited to fitting 1.5 wavelengths. Each level up is one more particle. Since each level up requires exactly the same amount of energy, each particle that is created in the field has exactly the same energy (read, mass).

So, when we were talking about first exciting the springs (the mass energy) before we can get a wave going in the field (see Sect. 6.2), we meant that the field must get excited to the level that it holds one particle! After this is done, the wave can get momentum by adding extra energy. That energy does not go into the springs, but into the rope (see Fig. 6.4).

The mass energy is really the energy that gets stored in the springs when the field gets excited. Consequently, this is called a field quantum. And it becomes clear why we are talking about “quantizing the field”. So “second quantization” shows us how a field can be quantized to hold “particles”, simply by showing how the mass energy of a particle can get stored in the field and showing that this mass energy is discrete: it is quantized.

So how does this work for massless particles? Are these field quanta too? Yes, they are. But where are the springs to store their energy in? Let’s dive into this issue. The thing is that the vacuum has a particular elasticity. So even when there are no springs attached, exciting a wave in a vacuum field still requires energy. However, this energy only consists of momentum energy since there are no springs attached. Since there are no springs, the wavelengths allowed are continuous and not discrete. So, one can excite field quanta in a massless field of any frequency one wants. The energy will be directly related to the wavelength by $E \sim F = C/\lambda$.

Once the field is excited with a certain amount of energy (i.e., frequency and wavelength), this energy will not change. So, the field does store energy in the elasticity of the vacuum. Creating a wave with a certain frequency requires an energy related to that frequency. One cannot use half that energy to create half a wave of that frequency. Hence, this wave is quantized too. The role of the springs is played by the elasticity of the vacuum and the energy is stored in that. So, in this case we can speak of a field quantum too. Each time one creates a wave with a certain frequency, it is a new field quantum. Each field quantum can only be absorbed as a whole. And so, we see the particle behaviour appear in massless fields such as light.

7.3 Phonons

When we look at media that are closer to home, e.g., water or a metal, these media can propagate waves too. Take a metal. It has a huge number of atoms that are attached to each other by a force. When one atom starts to oscillate, the force propagates the oscillation to the next atom, and that in turn gets excited, and so on. So, we may view the metal as a medium that can carry a field. For instance, a heat field: when we heat up one end of a metal stick, soon after, the other end becomes hot as well. So, the

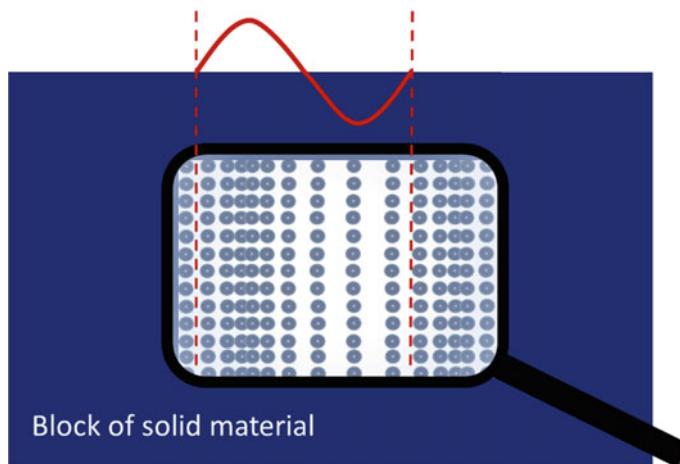


Fig. 7.6 If one could look at the individual atoms in a solid material using a magnifying glass, one could witness a phonon as a wave through the connected atoms in the material. The wave essentially consists of areas of packed atoms (high density) alternating with areas of fewer atoms (low density) moving swiftly through the material

atoms excited by the heat propagate that to the next atoms, and on and on, until the heat has reached the other end of the stick.

On a more delicate scale we could excite one atom, thereby creating a wave across the atoms with a certain energy and propagated through the material (see Fig. 7.6) until it deposits its energy, e.g., in a piece of measurement apparatus. We call such a wave a phonon, a term introduced by Igor Tamm in 1932. It behaves exactly like a field quantum, but the medium is different than the vacuum. Quantum field theory is used, e.g., in solid state physics (the physics that researches solid matter) to describe phonons in a solid. Similar characteristics of particles in the vacuum can be found in phonons as well. Hence, the same theory can be applied.

7.4 Conclusion

Waves in fields have a quantized energy. Massive waves store that quantized energy in the springs attached to the field. Such an excitation of the field is called a field quantum. The springs require it to have a specific discrete energy, the mass energy. It is this energy that creates the different dispersion related properties discussed before.

Massless waves store only momentum energy in the elasticity of the vacuum and can have any frequency. The field quantum associated with this has an energy that is fixed to that frequency.

Chapter 8

Energy in Waves and Fields



In the previous sections we were talking about the kinetic energy and the potential energy. We showed how a wave in a field is a continuous interplay of velocity and potential, like the pendulum. Let's dive a little more deeply into this.

A particle is assigned a particular energy and feels a particular potential. A field does not have one location, like a particle. So how can we assign a particular energy to a field? The only thing we can do is assign the energy the field has at each position and say what the strength and direction of the potential is at that point. Hence, we get to define the energy density and the potential density. This means that when a particle is represented by a group of waves, we can see the average energy density develop through the medium (see Fig. 8.1).

The total energy of the particle can be found by summing up its energy density at all positions in the field.

Classically, we could look at the potential energy of a particle and its kinetic energy (as with the pendulum). The way the potential energy gets exchanged for kinetic energy tells us how the particle will move.

In QFT we have to do the same, but based on the energy *density* and the potential *density*. The exchange of potential energy for kinetic energy and back will give a wave in the field.

Let's make this a bit more explicit: when we excite the field at one point, that point gets a high field amplitude. Since the vacuum is elastic, it will pull the field up at nearby points. This is the same as applying a potential to those nearby points. These points will get a velocity, go up themselves, and in turn provide a potential to the next points. And so on (see Fig. 8.2). So, when we work with an energy density and a potential density, we assign energy and potential to each point in the field. The elasticity of the vacuum relates these points to each other and so they react to each other. The result is waves propagating through the vacuum. So, working with potential and energy densities gives us waves in the vacuum.

So, the potential and energy densities determine how the field will move, or rather how it will wave. When the field behaves like a harmonic oscillator, there is a continuous exchange of kinetic energy to potential energy and back.

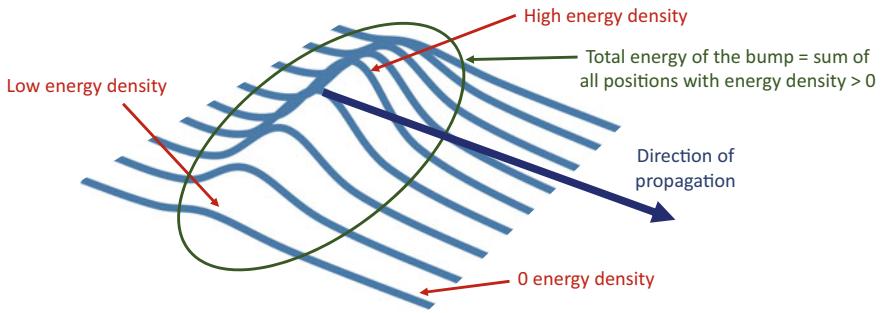


Fig. 8.1 Energy density distribution in the field

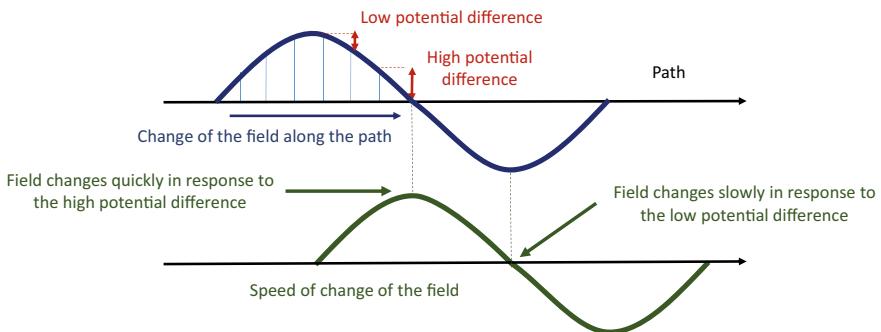


Fig. 8.2 Interplay between potential difference and kinetic energy of a wave in a field. When the field amplitude is at its maximum, the up-down movement of the field comes to a halt and reverses (it changes from an upward movement to a downward one): the potential is maximum, the potential difference between adjacent points in the field is minimal and there is hardly any kinetic energy (up/downward movement). Later, when the field amplitude is around 0, it is the other way around: no potential energy, high potential difference, and maximum kinetic energy (moving from upward to downward quickly)

So, everything now depends on the different types of potential the field will experience! As it turns out there are quite a few. In the following chapters we will meet them all, but before we do so let's first see if we can deepen our understanding of energy and momentum densities.

8.1 Conservation Laws

Now that we have to work with energy density and momentum density, how does this get transported properly? One can assign an energy and momentum to a particle and one can describe the *conservation* of these quantities. Energy and momentum can be

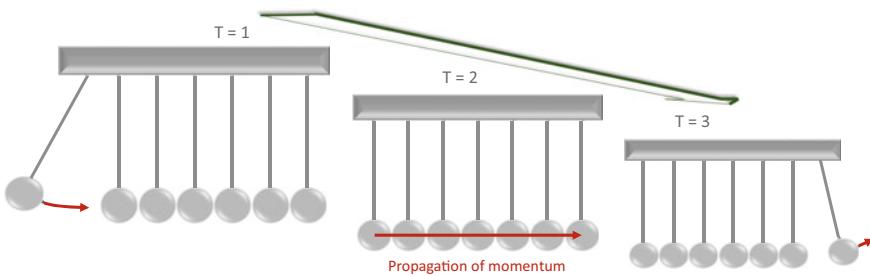


Fig. 8.3 A row of metal spheres passing on momentum

transferred, but will not (dis)appear in a magical way. Once one knows all sources of energy and momentum, it is simply a matter of bookkeeping what potential gets exchanged for what energy and vice versa.

So how does the bookkeeping work in a field? Let's first get a grip on how momentum and energy can move in the field. We will do this using the concept of flux. By flux we mean how a density moves from one place to the other. Suppose you have one of the well-known rows of metal spheres as shown in Fig. 8.3 and commonly called Newton's cradle. When you raise the leftmost sphere and let it fall back against the remaining, stationary spheres, after a short while the rightmost sphere gets an amplitude and kicks back the line. Then the leftmost sphere gets an amplitude. And so on. The momentum put in in the first step is propagated through the spheres. In a similar way, a field can pass on momentum from one place to the other via a wave.

The difference is that a wave can extend over space and can consist of a packet of waves, a bump, or a group. What we can do to measure that momentum and energy coming through is to create a sort of “screen” *in space* through which the wave passes. All the added-up momentum and energy that passes through the screen must be the total momentum/energy *flux* from A to B. We can place just such a screen *in time* to sum up all the energy and momentum *density* that was streaming through from moment 1 to moment 2 (see Fig. 8.4). In this case, we sum up all the energy and momentum available in a space at moment 1 and we do that sum again in space at moment 2 and find that the total has not changed: energy and momentum are conserved. Remember that energy is proportional to frequency. So, when we say that energy is conserved in time, what we are really saying is that the average summed up frequency in the fields involved will not change in time. Similarly, momentum is inversely proportional to the wavelength. So, conservation of momentum means that the average summed up wavelength will not change in time.

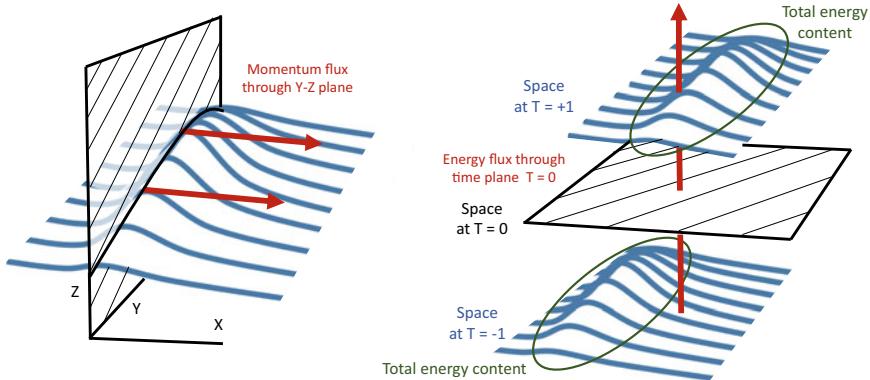


Fig. 8.4 Screen in space to add up momentum passing through (left) and a screen in time to add up energy passing through (right)

8.1.1 Energy—Momentum Tensor

The energy–momentum tensor contains all elements of energy and momentum density as well as energy and momentum flux. We can use it, e.g., to describe the laws of conservation. A tensor is a matrix that describes how vectors change size and direction. For instance, consider the wind in the atmosphere. At each point in the atmosphere the wind has a (slightly) different direction. Just look at weather charts to see that. So, you can assign a wind vector to each position in space. But these vectors change in time. And each of them may change differently. A tensor describes how the wind vectors change. In fields, the momentum density is different at each point in the field. We can assign a momentum vector to each point since the momentum has a different magnitude and direction at each point, just like the wind in the atmosphere. Energy only has a magnitude and no direction. For instance, the energy of a ball has no direction. The direction the ball is moving in is the direction of its momentum. However, in a field, energy can have a flux: it can flow (i.e., change) in a particular direction. The momentum passing through the screen in Fig. 8.4 is essentially the momentum flux in the x direction.

So, the density in fields can change in time and space. Suppose you have a cube with a certain momentum density in it at one point in time. In the direction of the momentum, the momentum can change (it will flow away). However, the momentum can also change in other directions, for instance in the case of a force changing the direction of the momentum, sometimes called shear stress. So, we can have a different inflow or outflow of momentum into or out of the cube in each direction (see Fig. 8.5).

This means that we have to define a different momentum flux vector for each direction the momentum has in space. When we have three momentum directions times the three directions of the momentum flux, we need a matrix of $3 \times 3 = 9$ numbers to describe the momentum flux for all three momentum components. When

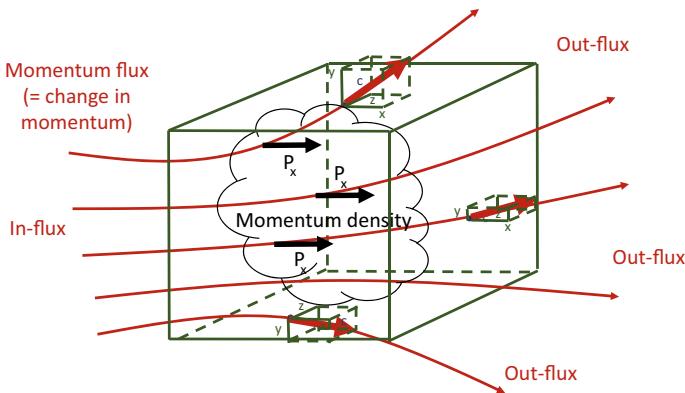


Fig. 8.5 The change in time of the momentum density is equal to the net inflow/outflux of momentum density. The same goes for the energy density. This is the same as saying that energy and momentum do not get lost (conservation law). It only gets moved from place to place and changed into other forms

we add the energy density and flux as a fourth component, we need $4 \times 4 = 16$ numbers (see Fig. 8.6).

The rows of the matrix tell us something about the conservation laws. The way the density (e.g., the density of momentum in the x-direction) changes in time equals the total inflow and outflow of this density in all three directions. That makes sense: when there is a net inflow of momentum, the momentum density grows over time. When there is a net outflow of momentum, the momentum density decreases over time. The same goes for the energy density. The growth in energy density equals the net amount of energy inflow, etc.

Flux through the plane $X = \text{constant}$			
Energy density	Energy flux in X direction	Energy flux in Y direction	Energy flux in Z direction
P_x density	Flux of P_x in X direction	Flux of P_x in Y direction	Flux of P_x in Z direction
P_y density	Flux of P_y in X direction	Flux of P_y in Y direction	Flux of P_y in Z direction
P_z density	Flux of P_z in X direction	Flux of P_z in Y direction	Flux of P_z in Z direction

Change in time of the P_x density = net in + out flux of P_x in all directions

Momentum pressure is equal to the net in + out flux in the direction of the momentum.

Fig. 8.6 Energy—momentum tensor

A column in the matrix tells us something about the flux going through a specific plane, for example, the total flux going through the plane that is defined by $x = \text{constant}$. Through this plane we find an energy flux, a flux of x -momentum, a flux of y -momentum, and a flux of z -momentum.

The conservation laws originate from Noether's theorem, after Emmy Noether, a mathematician who lived around 1900. She laid the mathematical groundwork for the definition of the conservation laws [Ref. 65]. In a non-mathematical way, Noether's theorem goes as follows [Ref. 66]:

If a system has a continuous symmetry property, then there are corresponding quantities whose values are conserved in time.

The propagation of these quantities in time is called a Noether current.

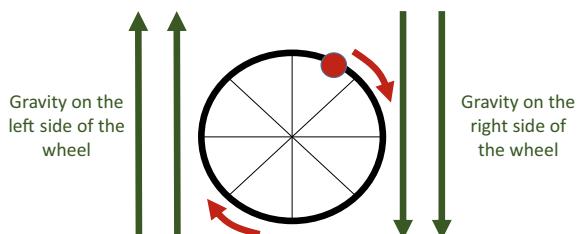
For example, the laws of physics do not change in different parts of space–time. The symmetry is that when you translate a system to a different part of space–time, the same laws must be valid. The theorem states that “there are corresponding quantities that are conserved in time”. One such quantity is energy.

For example, suppose that the laws that describe how energy is exchanged were different in one part of space compared to another. In that case you might be able to create energy out of nothing. For example, take a wheel with a mass on it. Gravity could pull the mass down on one side, making the wheel turn, and if gravity pointed upward instead of downward on the other side, the wheel could gain energy indefinitely, turning faster and faster (see Fig. 8.7). That would be a great source of energy!

The fact that the laws of physics do *not* change in different parts of space–time (or rather, the fact that they are everywhere the same continuous function of space–time) means that it is only possible to exchange energy under the same conditions everywhere. So, an increase in energy must come from a source that experiences a decrease. When the energy is transferred back, it must be done under the same conditions, leading to an equal decrease and an equal increase in the source. So, we are exactly back to the original situation. In that case, the wheel as described above is not possible. Nature turns out to be an excellent bookkeeper, simply by keeping the rules the same at all times. In this way, energy conservation follows from the symmetry that the rules do not change over time.

In the same way we can state that the bookkeeping of energy and momentum density is perfect since the laws of physics which govern their propagation do not

Fig. 8.7 The impossible energy pump



change in space and time. Hence, the momentum density flux and the energy flux are Noether currents: they are conserved under translations in space and time.

Another example is the conservation of orbital momentum. The laws of physics are symmetric with respect to a rotation in space and time (meaning that when you turn something around the laws of physics are still the same). Therefore, Noether theorem implies that the orbital momentum cannot change unless a force acts on it. A force provides a direction in space, e.g., you can use the force to determine a direction in line with the force, opposite the force, or perpendicular to the force. Hence, you may see the force as breaking the symmetry. But as long as there is no force, orbital momentum is conserved.

Now that we have some grip on energy and momentum densities and understand how they behave, let's look again at particles and how they can be represented by waves in a field. This time from the perspective of fields and densities.

8.2 How to Envision a Field Quantum

We have discussed before how a wave can be a particle, but now we have stretched the idea from ropes (1 space dimension) to fields (3 space dimensions) and density in a field. How is such a density a particle? When a particle is created, we said that energy must first be put into the springs attached to the field. These springs are quantized, i.e., one particle represents one energy and one frequency in the springs. For one spring, we showed how to fit a wave in that spring, resulting in the quantization of the energy levels in the spring. One energy level equals one particle.

That's great for one spring, but the whole (3D) field is full of springs. And these springs are connected via the elastic vacuum. So, there is never just one spring oscillating: multiple springs are in different stages of oscillation when a particle is in the field. These springs all oscillate with one frequency, but the energy contained in the field is more than we find in one spring. The number of springs that are in oscillation at one point in time depends on the elasticity of the vacuum. If the vacuum were completely inelastic, exciting one spring would mean that the vacuum would pull up a large number of other springs as well (see Fig. 8.8).

8.2.1 Coupled Oscillators

A single spring's quantization requirement relates the amplitude of the field to the mass wavelength (see Fig. 7.5). However, when there are springs attached, this attachment interferes with the way the whole thing moves. The springs could very well be in each other's way and the wave as a whole may not be sustainable. In that case, the springs would be transferring energy to one another and the wave would become unstable or get damped out.

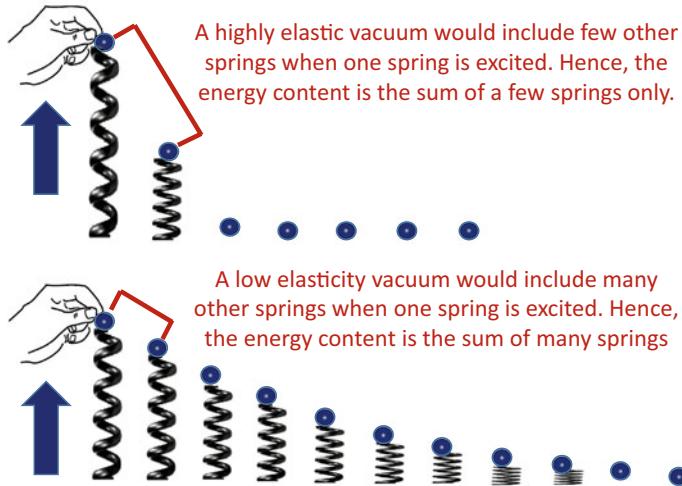


Fig. 8.8 Number of springs in oscillation when the vacuum is highly elastic (top) or when it is hardly elastic (bottom)

Let's look at a commonly used model to show how this works. In this model we have two masses and three springs attached to each other. A sustainable wave is characterized by the masses moving in such a way that they strengthen each other's motion. This means that they have to move in line or in exactly opposite directions. Indeed, for a model with two masses, these are the two ways they must oscillate: in line or opposite (see Fig. 8.9). The two ways relate to a particular wavelength of the wave as a whole. These particular movements are called modes or eigenfunctions of the system of coupled oscillators. Such a specific mode can be described as *one* harmonic oscillator. The easiest way to see this is by looking at the movement in line: the two masses act as if they are one harmonic oscillator. So, we can conclude that *a set of coupled harmonic oscillators can be considered as one harmonic oscillator*, but only when they are in a specific mode with a specific wavelength. If we had three weights connected by four springs, we would find three such modes. So, the number of modes we can find in a system is equal to the number of weights (in this example).

Let's look at another example: phonons. For simplicity's sake we take just 7 atoms in a crystal lattice. The atoms attract each other, so that when one moves up and down, it brings the nearby atoms along with it. This situation looks similar to our weights on a spring. It is also a coupled harmonic oscillator. So, what modes can we expect in this system? The lowest momentum mode is the longest wavelength we can fit on the 7 atoms. For the sake of argument, we take the end atoms to be fixed or to always move in line with each other (which is called a periodic boundary condition). The longest wavelength to fit on the 7 atoms is then half a wavelength across the 7 atoms.

The highest momentum mode corresponds to the lowest wavelength of twice the distance between the atoms. The reason for this is that shorter wavelengths cannot be supported. In the case of the shortest wavelength, each neighbouring atom moves

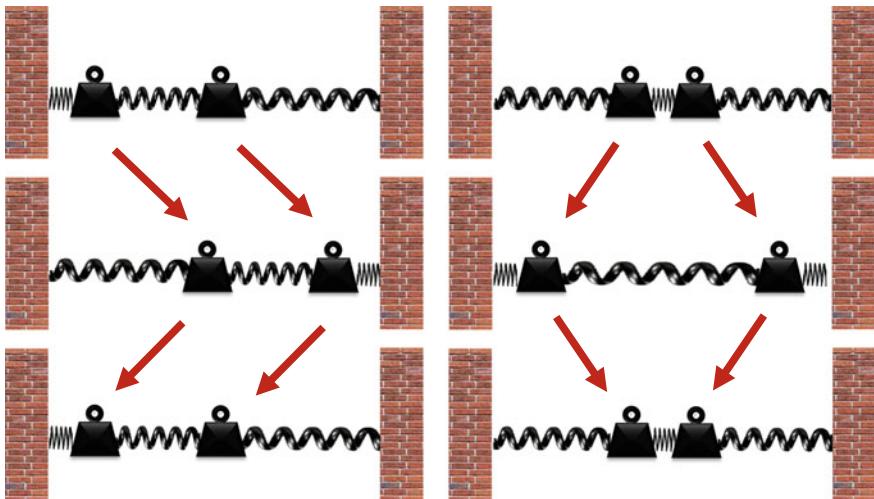


Fig. 8.9 There are two ways that two masses connected with three springs could oscillate in line. Each of these examples can be described as a harmonic oscillation on its own

in the opposite position (see Fig. 8.10). Try picturing a much shorter wavelength and ask yourself how the atoms could create that.

The result is that there are 6 possible modes. In general, the number of modes is equal to the number of atoms, but we fixed the end points and this reduces the number of modes by one. Hence, we see that phonons can accept energy only in discrete amounts: the amounts related to the available modes. So, a phonon is also a quantum and behaves much like fundamental particles in the vacuum. Any wave in the crystal can be described as a superposition of these modes or quanta.

The entire set of all possible phonons is called the phonon density of states which determines the heat capacity of the crystal. After all, the heat the crystal can store,

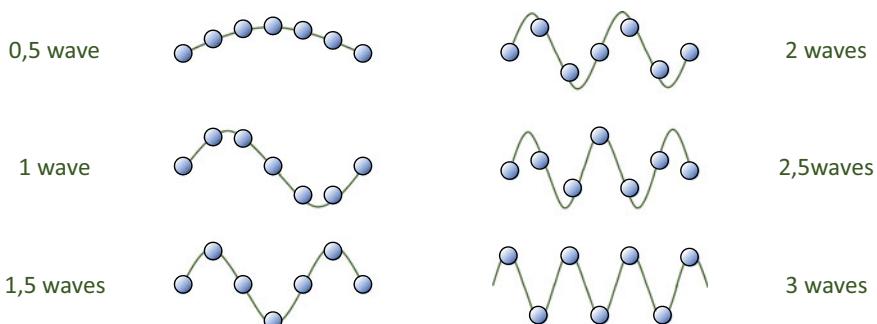


Fig. 8.10 Seven atoms with 0.5–3 waves. There are only 6 modes instead of 7 because the endpoints are either fixed, or always oscillating in phase (periodic boundary condition)

it can only store in the movement of the atoms, hence in the waves. At 0 K there are no phonons. At any higher temperature, there is a mean number of phonons in the crystal, depending on that temperature. The multitude of phonons behaves exactly the same as a box with black walls that contains photons. The heat inside the box is determined by the number of photons that get absorbed and emitted by the walls in thermal equilibrium. When we consider an infinite number of connected weights or atoms in a lattice or the like, the number of possible modes and wavelengths becomes endless.

The system we are considering for massive waves (the rope connected by springs) is of a slightly different nature. In this picture we have two types of “springs”: the “mass springs” and the elasticity of the vacuum (which we can view as a kind of spring). Now it becomes more complicated to view the way modes are created in this system. Nevertheless, there are also modes in this case. Since we usually talk about an endless set of springs, the number of modes is endless as well. Each such mode can be treated as a single harmonic oscillator with a specific frequency. When a massive field quantum (particle) is created, a mode is in fact created with a specific frequency. The frequency is determined by the strength of the springs (i.e., by the way the field interacts with the Higgs potential) and the momentum of the quantum. So different wavelengths and (group) velocities are possible. The quantum can change mode by speeding up or slowing down. Mathematically, when a quantum is created in the field a wave is created with a specific mode depending on the mass and momentum.

What does the connectedness of the springs do to the amplitude of the field? Just one disconnected spring would have the amplitude that fits with having all the energy stored in that one spring. This single spring’s oscillation can be quantized as in Fig. 7.5. But how does that work when more springs are attached? Basically, moving two springs at the same time would require twice the energy. So, if the energy of a quantum had to be spread over two springs, their amplitude would be smaller.

The more springs are involved, the lower their amplitude. The quantum wave would still fit an oscillator, but now this oscillator consists of a mode of many springs together, as in Fig. 8.10. The amplitude of the oscillator relates to the combined amplitude of all the springs involved.

When the quantum spreads over space, more springs get involved, and their average field amplitude decreases. The energy content (and hence the frequency) will still be the same, even though they oscillate together. The fact that the frequency remains the same can be understood to some extent by comparing with the pendulum: the amplitude of the pendulum does not impact the frequency. This is how a wave can spread over a large region of space, while the springs essentially remain “connected” into a combined total energy.

When a quantum gets a velocity as well, the frequency of the wave as a whole goes up. When this frequency goes up, so does the velocity, while the wavelength goes down. The quantum is free to spread over space and to spread the wavelengths and velocities of the waves it consists of, while maintaining the same mass frequency.

So far, we have been discussing the particle as a plane wave. As discussed before, we can describe the group of waves comprising a particle as one wave with a group velocity. Keep in mind though that, in reality, the quantum will consist of a number

of different waves and the energy will be spread over them. Each wave (each mode) has an amplitude, and the combined amplitudes make the quantum. Therefore, the quantum is really a wave packet.

In the mathematical description, a field is described as a sum of modes times an amplitude for each mode describing the number of excitations (quanta) of that mode. So, the total field amplitude is a sum of all modes multiplied by their amplitudes. This shows how a field can contain multiple quanta.

What happens when there is more than one quantum in the field? When there are two bosons in the same place and in the same state (i.e., the same mode), the amplitude of the field goes up to the amplitude that harbours two quanta (see Fig. 7.5) and the field's total energy content is $2hF$ instead of hF . If two such quanta reside in *different* positions, the amplitude around these two positions will be one quantum each. When they meet up, the total field amplitude will be forced to move up to containing two quanta at the position they meet. We will see later why this only works for bosons and explicitly not for fermions.

So clearly, both frequency and wavelength can change as a result of gaining velocity, leading to a massive quantum with a particular (group) velocity. So, a quantum is not a simple one frequency/one wavelength thing but rather a complicated set of waves with different wavelengths and frequencies. Hence, quantities such as energy density and momentum density make more sense for describing quanta in a field than the classical momentum and energy would.

8.3 Annihilation of a Field Quantum

The ideal field quantum is created as a mode in the field. This mode can be described as a single harmonic oscillator. This is called canonical quantization. Canonical refers to the absolute “eigenstate” the system can be in, which is orthogonal to the next eigenstate. Orthogonal means that the two eigenstates are completely independent and different states. As mentioned before, the reality may be different. What would happen if the quantum were not in a perfect mode? In this case the quantum would no longer be in a sustainable wave and the wave could fall apart. We are going to see different examples of this later when we look at absorptions of photons. In this case, the “canonical quantization” breaks down.

A pressing problem is, for example, how to view the annihilation of a quantum, when the quantum is spread over some distance in space. In this situation, how can it be annihilated at one position?

One way to view this problem is by considering that the energy content is related to the frequency of the springs. Suppose that a bit of the waves gets absorbed at one position. Then the energy content of the waves may drop below the minimal energy content of a quantum (i.e., below the energy level of the quantum at 0 velocity). So, what would happen? Remember from Fig. 7.5 that an ideal quantum is a wave that has an energy level at which it exactly fits within the potential walls of the Higgs field it interacts with. Even when the quantum is not perfect and built up from

multiple waves, these waves need to add up to exactly that energy level. When some of the energy gets absorbed, the energy in the combined waves is not enough to fit exactly within the potential walls (see Fig. 8.11). When some energy is absorbed at one particular spot of the wave, it means that the spring at that point will no longer oscillate as much. This means that it will drag down the adjacent springs. The consequence for the wave as a whole is that all the springs together will not have enough energy to reach the potential walls.

The result is that the mass wave will start to interfere negatively with itself. To understand this, picture a wave (see Fig. 7.1) that is not 0 at the potential wall. It would reflect back against the wall and start to interfere negatively with itself. So, the oscillation of the springs will start to die out. This happens to all connected springs, and so the quantum ceases to exist. In the process, the energy needs to go somewhere, and the absorbing entity has no other choice but to absorb all of it. Or in the case of a particle annihilating against its anti-particle, the energy is freed to produce, e.g., a high energy photon.

The disappearance of the wave is one way in which the wave function can “collapse”. The energy is no longer available for others to be absorbed and the field quantum can no longer be maintained.

Another form of “collapse of the wave function” happens when a field quantum that is spread across space gets measured at one particular position. We will get back to this in Chap. 12 when we discuss quantum decoherence.

8.4 Describing a Field Quantum

We have discussed how a field quantum consists of a sum of many waves, each of which is a mode of the field. That is a complicated picture. We have also seen that coupled harmonic oscillators can be described in a simple way as one harmonic oscillator. That is a great simplification. The same is done in field theory, where a quantum is described as a plane wave. From an energy perspective, it is also described as a plane wave continuously exchanging the potential energy of its bare field with kinetic energy, as depicted in Fig. 8.2. That does not mean that we do not sometimes have to return to the picture that, in reality, a quantum is built from many waves in a wave packet.

In this bare wave we only encounter the elasticity of the field. Since this is the elasticity of the vacuum, each description starts with a massless wave in the bare field as if nothing else were there: no Higgs, no nothing. The motion of the field depends only on its initial excitation and the potential delivered by its elasticity.

So now that we have simplified the picture of a field quantum to that of a single bare plane wave, we can let the fun start. A quantum does not exist on its own. It will interact with many other fields, so it will have to deal with all sorts of potentials. In quantum field theory, a quantum is described by all the potentials it has to deal with. All the potentials together will in the end tell us how the quantum will move and behave. Just as we use the gravitational potential to tell us how a ball will fall.

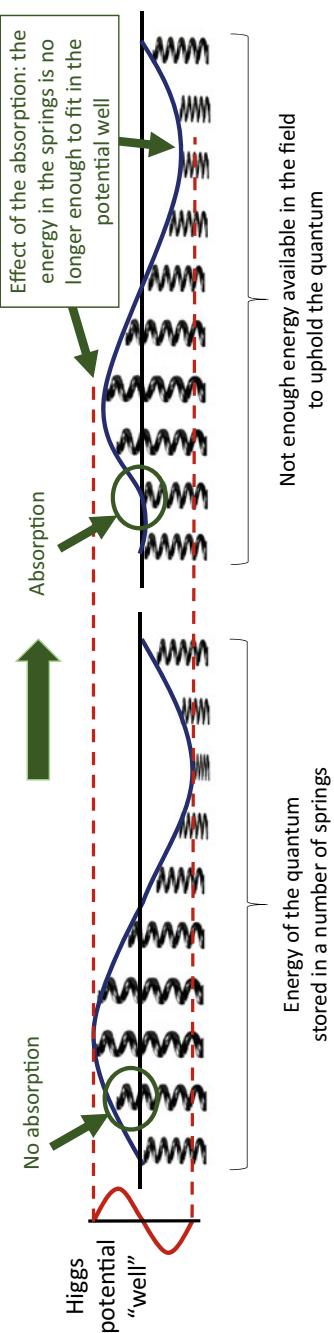


Fig. 8.11 Annihilation of the wave function. When some energy is absorbed anywhere in the wave, the total energy can drop below the point where the springs can oscillate between the potential walls and they will no longer be able to do so. The springs cannot sustain the wave and the oscillation dies out completely

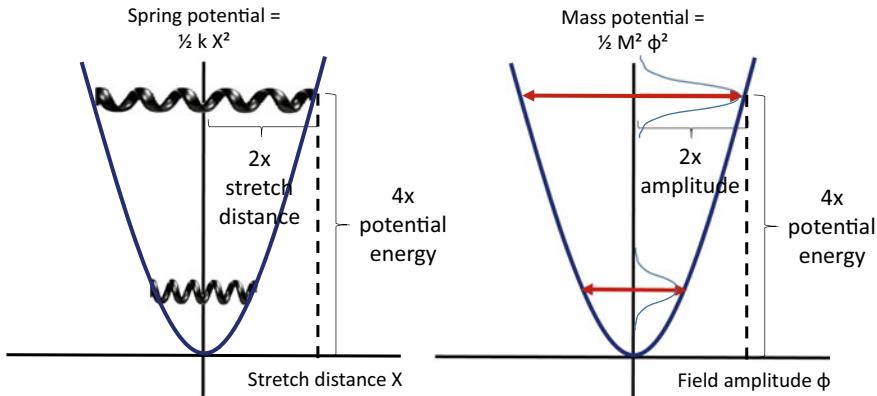


Fig. 8.12 The mass potential compared to Hooke's law

Each of these potentials has its own effect. One of them we already met: the Higgs potential.

Often, the Higgs potential is described as a “mass potential”. The mass potential (or mass spring) shows exactly the same behaviour as a spring in Hooke's law: when the field amplitude is twice as high, the energy it requires is four times as high (see Fig. 8.12).

When we compare with the mass potential, we see that the field amplitude Φ can be viewed as corresponding to the stretch distance X of a spring. Remember that the field amplitude is the amplitude of a wave in the field. Furthermore, the mass² then corresponds to the spring strength. This means that, as in Fig. 7.3, doubling the mass² will result in doubling the potential energy when the field is stretched to the same amplitude. So, in fact, the mass potential that a field quantum experiences has exactly the same form as that of a spring. This is why we can use the model of ropes and springs so effectively. We know the effect of this potential by now: it acts like the springs attached to the rope and creates the mass-like behaviour of the quantum. So, this potential is added to the particle wave to make the bare wave massive and behave as such.

In the following chapters we will investigate other potentials that play a role. There is an important category of potentials that relate closely to symmetries. We will see that this category is responsible for forces, so we will pick up from there in the next chapter.

Chapter 9

Symmetry and the Origin of Force



9.1 Rotational Symmetry

Symmetries play an important role in quantum field theory. There are different types of symmetry, and these are all described mathematically by group theory.

What is a symmetry? When you look at your face in the mirror, the left side of your face looks like the mirror image of the right side: your face is symmetric. A plain ball is symmetric in the same way: the left side is the mirror image of the right side. However, there are more symmetries. That same ball is, for instance, circularly symmetric: if you turn it around its center it looks the same. It does not matter how far you turn it around, it looks the same.

So apparently, something is symmetric when it looks the same after some operation. An operation can be “take the mirror image”, or “turn it around”, or “translate it over a distance in space”. A set of operations is called a symmetry group of an object when it comprises all the operations leaving the object invariant (indistinguishable). For instance, for an equilateral triangle, the operation “turn through 120° ” changes the triangle into an indistinguishable triangle. The operation “turn through 240° ” does that as well, just like “turn through 360° ”. All these operations are members of the symmetry group of an equilateral triangle.

9.1.1 Rotations in a Plane

Let's look at rotations in a plane, and more specifically, only those rotations that fix a circle with radius 1. Suppose you have a vector that points from the center to one position on the circle. Now suppose you turn it around through an angle. Then you still point to a position on the circle. So, a rotation of such a vector through any angle is a symmetry operation. The group of such rotations is called $U(1)$, the group of unitary rotations in a plane. Unitary means that the rotating vector has length 1 so it describes rotations that remain on a circle with radius 1. Consequently, you can apply

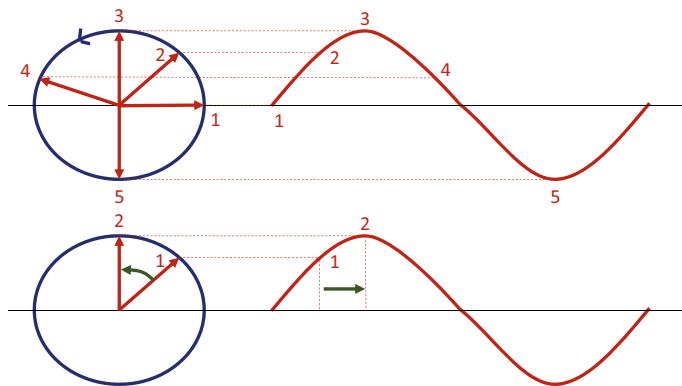


Fig. 9.1 Wave described as a vector in a plane. A rotating vector in a plane can describe a wave as follows: take the angle the vector makes with the x axis and put that angle on the x axis in the graph on the right. Then take the projection of the vector on the y axis and copy that on the y axis in the second graph (the red dotted lines). The result is the wave on the right

any rotation as many times as you like and you will stay on the circle. Operating on a vector with a U(1) rotation would imply that you change the angle of the vector but not its length.

You can use a vector rotating around the unit circle to describe the phase of a wave (see Fig. 9.1). The phase is one-dimensional. It seems as though you need two dimensions (the plane of the circle), but the radius of the circle in the plane is kept at one value, the unit value. Consequently, the only thing that matters in describing the phase is the angle. One variable (the angle) means one dimension. The wave function that describes the wave then assigns a value to each phase, the field strength (or in quantum mechanics the probability amplitude).

The 1 in the U(1) group means that we are talking about a group with a 1-dimensional operation (it operates only on the angle). Suppose you have a nice smooth wave. But then you operate on it with a U(1) operation. This means you shift the phase from one value to another. This is like rotating the circle in Fig. 9.1 through an angle (lower picture). The result is that you change the phase of the wave from position 1 to position 2. Since the wave function assigns a field strength to the phase, changing the phase implies a sudden change in field value (see Fig. 9.2). However, when the same phase change takes place in the entire universe, nothing

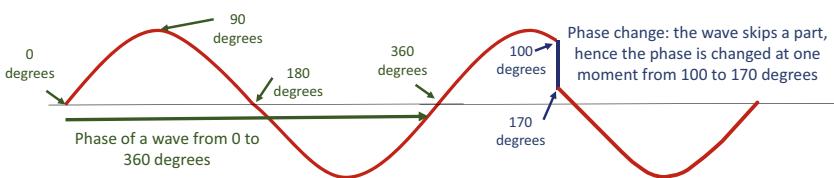


Fig. 9.2 A U(1) operation describes a phase change (blue) in the wave function (red)

really changes. It would be the same as taking the entire universe and place it 100 m further. Since all relative distances remain the same, nobody notices any change.

You might argue that a shift in phase changes the field value of some waves up and of other waves down (as in Fig. 9.2), since different waves will be in a different phase when the shift took place. So, all of a sudden, the relative field strengths will differ. This could be a problem if the relative field strengths of two waves at any given moment actually made a measurable difference. But before the phase shift happened, all waves were smooth. So, they will still be smooth after a global phase change. Furthermore, When the phase change happens throughout space at the same moment, none of the waves will change their energy (frequency) and momentum (wavelength). So, nothing really changes when we apply a phase shift.

The result is that a $U(1)$ operation leaves the world unchanged and is therefore a symmetry of the universe. In the next section we will see that nothing is as it seems in our wavy quantum world. We will discover how this symmetry is related to the electromagnetic force. But first, let's investigate some more symmetries.

9.1.2 Rotations in Three Dimensions

Some wave functions can change into each other. This means that at one moment you are a wave in a field and the next you are another wave in the same field. Or better, the wave has two “states” in the same field. An example is spin (as we will see later). A particle can be in one spin state (the particle “turns around” in one direction) and the next moment it is in the other spin state (the particle “turns around” in the other direction). The ability to change state is called an “internal degree of freedom”. You can also call it a symmetry. After all, when a particle can change state at will, it is like a symmetry operation that takes the particle from one state to the other or back. Just like rotating an equilateral triangle through 120° and back. When you rotate the triangle, it is in a different state (rotated) but that state is indistinguishable from the original state: the triangle looks exactly the same after the rotation. Hence, the rotation is a symmetry operation and the ability to rotate the triangle through 120° and back all the time is a degree of freedom.

So, what happens when you apply a magnetic field to, e.g., a fermion with spin up (i.e., turning around in one direction)? A magnetic field interacts with the spin of a charged particle. The consequence is that, while spin up and spin down did not previously show a difference, now they do: spin up gets a different energy level in the magnetic field than spin down. So, it will take energy to flip from one spin state to the other. It is no longer a symmetry. We say that the symmetry is broken and these two states of the fermion can no longer be turned into each other in this way. It requires energy to change spin down into spin up and these two waves are no longer symmetric.

Another example is that some particle types are actually two different manifestations of one and the same particle. These two manifestations can then be changed into each other at will. Again, this is a symmetry between two states and the ability

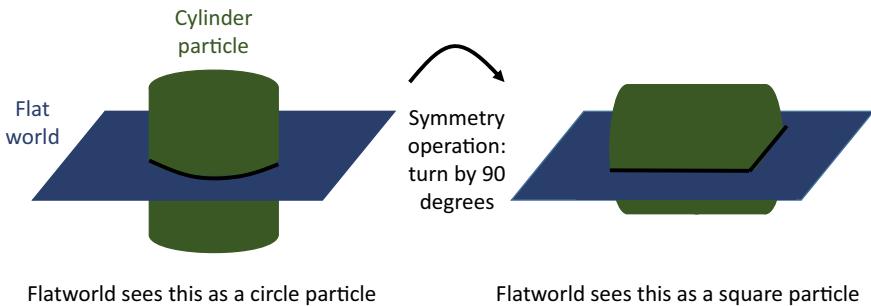


Fig. 9.3 How a cylinder particle can be two different particles in flatworld, related by the symmetry operation of a rotation in the third dimension

to switch at will is an internal degree of freedom. This may seem a bit abstract. Suppose you have a very flexible face (maybe you are the Hulk, John Travolta in “face off”, Ethan Hunt, or any other face changing movie character). Then you have the freedom to switch from one face to another as you please. That clearly is a degree of freedom. It is also a symmetry in the sense that both faces represent the same person underneath.

Another way of looking at this is by considering a particle to have properties in a dimension outside our four dimensions of space–time. To explain this, take an example from flatworld: the world of flat space in two dimensions. Suppose we have a particle that is a three-dimensional cylinder. When it resides in flatworld, the inhabitants of that world can only see two of the dimensions of the cylinder particle (see Fig. 9.3). So, for instance, they may see a circle-like particle. But when this particle is rotated through an angle of 90° , the particle suddenly looks like a square in flatworld! So, the people in flatworld might say these were two different particles, while all it takes is a rotation in another dimension to change one particle into the other. This is how we can understand what is called an SU(2) symmetry: it describes how one particle can change into another by defining a symmetry operation that “turns” one particle into the other. Obviously, both particles need to belong to the same symmetry group if it is to be possible to turn one into the other. The specific symmetry will only be able to turn these two particles into each other and not into other particles.

How is such a symmetry described? We can characterize a wave by two dimensions: the phase of the wave (horizontal axis in Fig. 9.1) and the field strength (vertical axis in Fig. 9.1). So, when we can change from one state (one wave) to another, this means that we need a symmetry operation that rotates a wave in two dimensions into another wave that also exists in two dimensions. This is a rotation in four dimensions.

At the same time the rotation cannot change the overall size of the wave. That is, the total summed up field strength of the first wave cannot be changed in the second wave, otherwise the total energy stored in the first wave would change when flipping to the other wave. So, we need a symmetry operation in four dimensions that does not alter the overall size of the wave. This is like a U(1) operation in the

sense that the size of the vector cannot change, but now in 4 dimensions. Such a symmetry is called an SU(2) symmetry. This means a “special unitary operation in two dimensions”. Two dimensions? But we needed four! In this case, the SU(2) is defined in two *complex* dimensions. It would take us too far to dive into complex numbers. To keep things short, a complex number has two dimensions of its own, so when we consider two complex dimensions, we are actually talking about $2 \times 2 = 4$ dimensions in the real space we are familiar with. The “unitary” in SU(2) means that the operation does not change size, just like U(1). The “special” means that the four dimensions are related to each other in a special way.

How to view that? When we consider rotations in four dimensions, we can find four ways to make a rotation, one around each axis of the four dimensions. So how can we find four rotations between two states? We can turn state 1 into state 2 and we can turn state 2 into state 1. These are two types of rotation. What other rotations can we think of? We can consider “self-rotations”: a state rotating into itself. After all, when you rotate a triangle through 360° (an allowed symmetry operation), you get back the triangle in its original state. So, we can turn state 1 into state 1 and we can turn state 2 into state 2. These are the other two rotations. However, there is a problem. The two self-rotations are related. We have to consider them as the same type of rotation. The result is that we have only three rotations: from 1 to 2, from 2 to 1, and one type of rotation of each state to itself. This means that we cannot use just any U(2) symmetry operation, since then we would be able to distinguish all four rotations. We have to use a special version of U(2) that is restricted to three rotations. This is called SU(2).

Later we will discuss how this symmetry is related to the weak force.

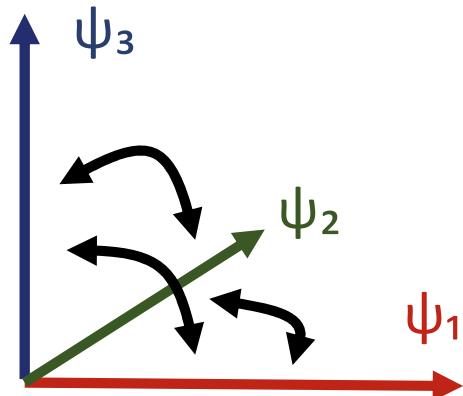
9.1.3 Rotations in Eight Dimensions

Essentially, the story is the same as with four dimensions, except that now we do not have *two* states that can be rotated into each other at will. We have *three* states that can be rotated into each other at will, and three waves, each with two dimensions, makes $3^2 = 9$ dimensions.

If we assign one dimension to each wave Ψ , we can represent this symmetry as in Fig. 9.4.

The nine dimensions suggest that there will be nine possible rotations. We should have rotations from 1 to 2, from 2 to 3, and from 3 to 1. Then we should have the same three rotations in the opposite sense. That leaves us with three self-rotations. However, in Fig. 9.4, how can the red arrow make a 360° turn? We can do that in two ways: via the blue dimension or via the green dimension. The self-rotations are related, just as in the case of SU(2). But now we have two possible self-rotations instead of three. That means that we have in fact eight dimensions instead of nine. Hence, this symmetry, called SU(3), is eight-dimensional. It is the special version of U(3), which is nine-dimensional. It describes how three waves can look alike and be rotated into each other.

Fig. 9.4 Rotations between three wave functions



An example are quarks and gluons. Quarks are the particles that protons and neutrons are made of, for instance. Gluons are the particles that keep quarks together, e.g., in a proton. They are responsible for the strong force. We will get to that later. Each quark and each gluon can be in one of three states. Each state is represented by a colour (red, green, or blue). Of course, they do not actually have a colour, but this is a way to describe the three states they can be in. They can change state at will. Hence, this is an internal symmetry or an internal degree of freedom. You can compare it with a Hulk that now has three faces to choose from. It can change from rosy red to green (as you know from the movies), but now also into blue.

This example shows that SU(3) is related to the strong force.

9.2 The Electromagnetic Field

The first symmetry we will look at is the U(1) symmetry. We will come back to the others later. We explained that a U(1) operation leaves the world unchanged and is therefore a symmetry of the universe. But I also promised that nothing is as it seems in our wavy quantum world.

You see, the problem is this: if we just take a wave in the field (massive or not), it is symmetric with respect to a global phase change. If we apply a phase change throughout the universe in one go, we will find that nothing measurable has changed. *But when the speed of light is the maximum velocity, how can we change the phase in the entire universe in one go?* The vacuum must propagate the phase shift and cannot do that in any way faster than its elasticity permits. But the propagation of such a phase shift is not described in the waves and potentials we have so far. So, something is missing.

Let's first consider what we are dealing with here. What happens when a phase shift takes place? It leads to a sudden change in the field strength. This "shock to the system" was supposed to be everywhere at the same time. But since this is

impossible, we must now see how to propagate such a phase shift (see Fig. 9.5). You could compare the phase shift to throwing a stone in the water. It leads to a wave propagating outwards. The phase shift will do the same. The phase shift leads to a phase shift a little further in space. That phase shift will lead to a phase shift still further down. And so on at the speed of light.

Why at the speed of light? When you throw a stone in the water, an excitation is created in a field. So, the phase shift creates an excitation in a field. At this point of our reasoning, it is not clear what field is excited. Let's assume for now that this is a field that does not connect to Higgs and is therefore massless. Then the propagation of the phase shift will occur at light speed.

Suppose the phase shift happens in an electron field, i.e., the field that contains an electron wave when it gets excited. The phase shift itself does create a wave, but it cannot be an electron wave. If it were, an electron would be spraying out electrons all the time (with each phase shift a new electron would be created). Clearly, there is not enough energy to do such a thing. So, it must be a different type of wave that results from a phase shift. But the electron field itself cannot harbour another type of wave. It can only contain excitations that are electron quanta. *So, a phase shift in a field does not create an excitation in that same field.* Consequently, the phase shift must be propagated in a *different field*, a field that gets created specially to propagate that phase shift.

Summarizing, the problem is that a global symmetry is not possible due to relativity, or due to the elasticity of the vacuum. A symmetry operation needs to be global. When it cannot be global (in one go for the entire universe), it must be propagated. Hence, we say that the symmetry can only be a local symmetry. Therefore, we must add a field that is able to propagate the symmetry operation in line with the elasticity of the vacuum.

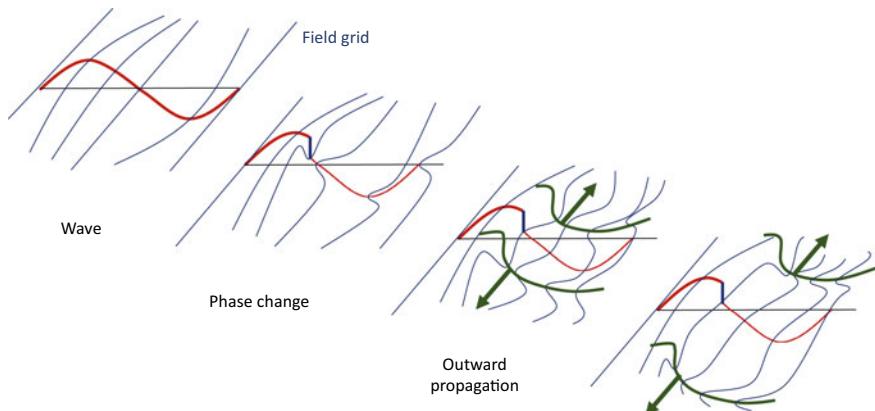


Fig. 9.5 Propagating a phase shift

The field that is generally added to make a symmetry local is called a gauge field. The excitations in such a field are called gauge bosons, since they generally turn out to be bosons. We will treat the properties of bosons later.

How can we describe such a gauge field? When a phase shift happens it creates a difference with the neighbouring part of the field that has not been shifted yet. *This difference is in essence a potential difference*. That potential difference gets propagated away by the gauge field. It is important that the field that produces the phase shift is “connected” to the gauge field. Being connected means that the potential difference of the phase shift is picked up by the gauge field. When it is picked up, it is the same as lifting up a rope in the gauge field. This sets off a wave in the gauge field (see Fig. 9.6).

Such a wave has all the things we expect from a wave in a field: it has a basic wave expression that describes the dynamics of the wave, an oscillation between potential difference and movement in the field. It can have a mass potential if the gauge field is connected to Higgs. We will see later why the U(1) gauge field is not connected to Higgs (see Chap. 17).

The potential it propagates is called the coupling potential or interaction potential. Basically, this potential (or spring) describes the connectedness of the field whose phase was shifted and the gauge field that propagates the shift away. When the potential is felt by another field, it looks like Hooke’s law again (see Fig. 9.7)!

In this potential, q describes the connectedness. It is also called the coupling constant. A is the potential (or field strength) of the gauge field and Φ is the field strength of the field that created the phase shift. This looks logical: the energy exchange between the field and the gauge field depends on their field strengths as well as the connectedness between the fields. Therefore, the spring strength is equal to $2qA$ in this case. So, the strength of the spring that is felt by the field φ is determined by the coupling constant q and the amplitude A of the gauge field. This

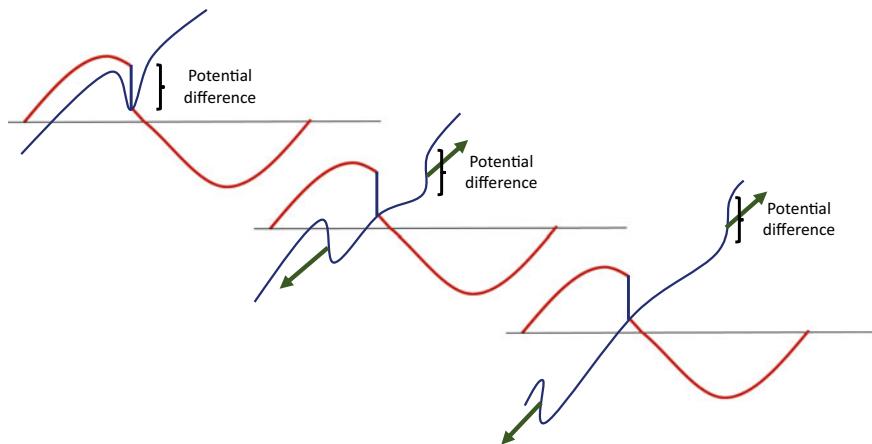


Fig. 9.6 A gauge wave as a result of a phase shift

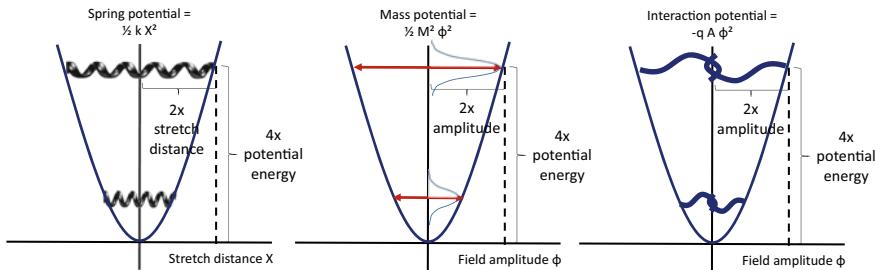


Fig. 9.7 The coupling or interaction potential has similar properties to Hooke’s law and the mass potential. The interaction potential can be seen as behaving like a spring

determines the potential energy that is felt by the field φ . Hence, the stronger the coupling, the more strongly the potential is felt; and the stronger the gauge field, the more strongly the potential is felt. This potential describes how the wave gets set off. It describes how the gauge field gets excited as a consequence of the phase shift.

This potential also tells us something else: it tells us what happens when a propagating phase shift meets another field quantum, e.g., if it meets a proton quantum in a proton field. Then this potential creates a “spring” that is felt by the proton, since the proton field is connected to the gauge field. The potential felt is a combination of both field strengths and their connectedness. So here we see that the basic features of a force appear! We will get back to how the force mechanism works in Sect. 9.4. We will show how the result of this process is a change in the velocity and direction of the quanta involved, and how “attraction” and “repulsion” work as a consequence of this.

Mathematically, when this is worked out for the U(1) symmetry we get an interesting result. *The basic wave and the coupling potential can describe the electromagnetic field!* Not only does this fit well with the field descriptions of the EM field, but it is also the simplest fit. Electromagnetism is the simplest gauge theory [Ref. 8, p. 129]. So, the electromagnetic field could actually be the gauge field that is needed to make the phase shift symmetry local. Hence, in quantum electrodynamics it is assumed that it is. Consequently, when a charged particle shifts phase, it creates a photon.

Taking this further, you could say that, if the speed of light were infinite, the phase shift symmetry would actually be a global symmetry and the gauge field needed to propagate the phase shift would not exist, i.e., the electromagnetic field would not exist if the speed of light were infinite. Consequently, you could say that the electromagnetic field is a direct result of the elasticity of the vacuum!

Now we can recognize the variables in the coupling potential: q is the electric charge and A is the electromagnetic potential. You may be familiar with the electric field E and the magnetic field B in electromagnetic theory. However, E and B are linked. An electromagnetic wave is a continuous oscillation between E and B . It turns out that E and B can also be described by one potential, generally denoted by A .

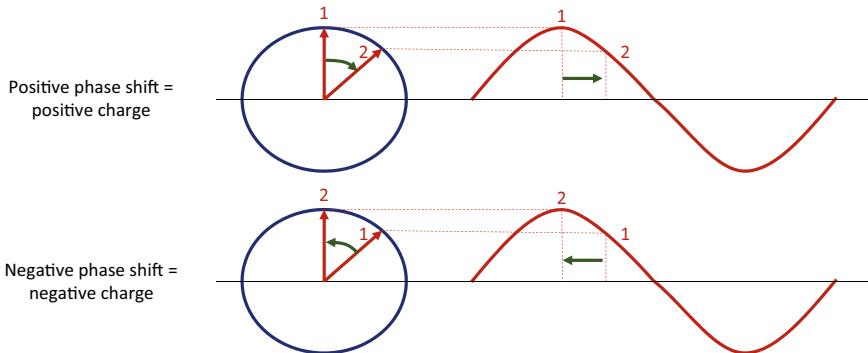


Fig. 9.8 Positive and negative charge are just phase shifts in opposite directions

A phase shift can take place in two directions. In Fig. 9.2 the phase changes from 100 to 170°. It changes forward, or in Fig. 9.1, to the right around the circle (see Fig. 9.8). It could also change backward, or in Fig. 9.8, to the left around the circle. A phase shift going forward (to the right around the circle) is associated with a positive charge, while a phase shift in the opposite direction is associated with a negative charge.

So, with q the electric charge, it is clear what its role is: the “charge of a particle” determines whether and how it feels the electromagnetic field. The charge determines whether and how a wave creates or experiences a phase shift. The way a wave (a quantum, a particle) creates a wave in the gauge field when it shifts phase is exactly the same as the way it experiences a propagating phase shift from elsewhere!

We said before that a global phase shift does not change the energy or momentum. This is the reason it is a symmetry in the first place. Clearly, when this symmetry can only be local, this changes: electromagnetic waves do carry energy and momentum.

The gauge wave can store information in the wave about energy (average frequency), momentum (average wavelength), and the potential it propagates away (the phase change). It spreads throughout space–time and so it carries information about position in space–time. It has a spin orientation, and so it carries spin information. There is no room for other information to carry in the wave. So, it cannot carry a charge. A charge (the ability to produce and feel phase shifts) would require the wave to hold information about the type of charge (negative or positive). The photon holds only one number besides the regular space–time, spin, momentum, and energy information one finds in a wave. This number is the phase change, a potential. The gauge group U(1) is Abelian. This means that it does not matter in what order phase changes take place, the result is always the same. Put differently, the photon has no (hidden) memory of the way it was created by the phase shift. It does not know whether it is the first phase change, the second, or the third, nor what phase changes have come before.

Table 9.1 Two types of fields in QED

Field	Wave type	Mass “spring”	Interaction “spring”
Charged fermion	Fermion	Yes	With photon
Photon	Gauge boson	No	With charged fermion

9.2.1 QED

When we put together the waves and potentials we have met so far, we get a landmark equation: the equation that describes the waves and potentials in quantum electrodynamics (QED). It tells us how charged particles move in an electromagnetic field. This is a very well tested theory and extremely successful: the measurements agree with the theory to sometimes extremely great accuracy [e.g., Ref. 60, p. 196; Ref. 30, p. 641]. In the equation we find two types of fields and it describes the waves and potentials for those fields (see Table 9.1).

So, there are two waves and two potentials in this equation. The potential that creates the mass of the charged particle (mass “spring”) and the potential that is propagated by the gauge wave and felt by the charged particle when it interacts with the gauge wave (interaction “spring”).

The interaction potential is meant to change the wave of the charged particle (that’s what an interaction does). So, the charged particle may be a perfect harmonic oscillator on its own, but with the interaction potential included, that will change. We will investigate further what this change looks like when we discuss virtual particles.

9.2.2 The Electromagnetic Field

So, we have seen that a phase shift in, e.g., an electron field leads to a wave that propagates out a potential. However, we also saw that the new wave is not an electron. Then we concluded that this gauge wave is a wave of the electromagnetic field. So, a phase shift in one field causes a wave in a different field. How should we view that?

Let’s go back to the example of fields in the atmosphere. Different fields in the atmosphere are the temperature field (a scalar field) and the wind field (a vector field). These are different fields in the same “carrier”, the atmosphere. As it turns out these fields influence each other too: when there is a substantial temperature difference in the atmosphere, this can lead to winds. Sometimes a particular type of temperature difference can even cause a tornado (this also requires the air to contain moisture, another field one can define in the atmosphere). So apparently, fields can set off behaviour in other fields in the same carrier.

The electron field and the electromagnetic field are not comparable to the fields in the atmosphere, but a similar principle applies. The electron field is a field that

is attached to the “springs” of the Higgs field. The electromagnetic field is not. So, the phase shift that leads to a potential difference in the electron field is propagated away by a field with very different characteristics.

Some characteristics of the electromagnetic (EM) field are:

The EM field is massless

The EM field does not couple to the Higgs field. So, it does not experience a Higgs “potential”: no Higgs “springs” are attached to the field. The field does not experience other potentials either. Hence, a wave in the EM field purely feels the elasticity of the vacuum and nothing else. That elasticity alone sets the velocity of waves in the vacuum to C, the speed of light.

Charge as a winding number

Mathematically, the phase shift represents electric charge. Charge is an integer that indicates the number of phase changes that the field naturally undergoes. We do not know why some fields undergo phase changes and others do not, but they appear to do so according to a fixed (quantized) number q. For most fields, q is -1, 0, or +1. However, +2 can also exist, for example. A special case are the quarks, which have a charge that is a multiple of 1/3, but they only appear in combinations that add up to integer values.

Sometimes the charge is referred to as a “winding number”, i.e., the number of times the EM field gets “wound up” in a phase shift.

Gauge invariance

Mathematically, the EM field turns out to be symmetric under local phase changes. This means that a phase change in the EM field itself does not lead to an extra EM wave. Put differently, it does not have a charge of its own. It does not couple to itself and it does not set off other EM waves. It has a charge “winding number” of 0.

We will see later that the gluon field does carry its own charge. This has a major impact on the behaviour of the field. The fact that the EM field does not produce EM waves on its own means that the influence of the field degrades over distance simply according to $1/r^2$. This would be expected when the only cause for degrading the field would be that it gets spread out over a larger sphere at a greater distance. The sphere increases its surface area by a factor of r^2 , so the field getting spread out must degrade by a factor of $1/r^2$.

Photons

The EM field can be second quantized. This means that the modes we find in the EM field can be considered as harmonic oscillators with discrete energy differences. Each energy difference is equal to

$$E = hF$$

So, the harmonic oscillator that can be related to a mode with frequency F has an energy difference of hF with the next level. This means that a wave in the EM field

can be created with a frequency F if we put in the energy hF . This is the first energy level above the ground state in the harmonic oscillator that is associated with this frequency. If we want to create a second wave, it takes another amount of energy hF . So, the EM field is quantized and the quanta have an energy level that is frequency dependent. Note how this is different from, e.g., an electron field, where the energy difference for creating an electron is determined by its rest mass (i.e., the Higgs springs) and is therefore a constant. It is not frequency dependent. There is only one type of electron with one (rest) energy.

This also implies that EM waves can only be created and absorbed in such quanta. These quanta are called photons. The mode expansion that results from the second quantization process allows two polarizations for each photon. These polarizations are referred to as spin +1 and spin -1, as if the wave can spin around like a helix. We will get back to spin later.

For now, it is interesting to address one aspect. The spin of the photon is spin 1, which means that it should be able to have spin values -1, 0, and +1. Spin 0, however, is not allowed. One can view spin 1 and -1 as “turning” around an axis that is parallel to the direction of motion of the photon (see Fig. 9.9). Spin +1 and spin -1 then involve turning in opposite directions around the same axis. Spin 0 would be perpendicular to the direction of motion. This would mean that parts of the wave must go faster than c , and other parts must go more slowly than c . This is not possible, so spin 0 is not allowed.

We can conclude that a phase shift in a field such as the electron field excites the EM field to produce a photon. The photon requires an energy hF to be created. The energy must be provided by the potential difference in the electron wave when the phase shift happens. This does mean that the electron wave loses energy. This is what happens when electrons get accelerated. They produce photons when they get accelerated and consequently lose energy in the process. As a result, they do not accelerate as much as one would expect. Part of the energy put into the acceleration gets lost again in producing photons.

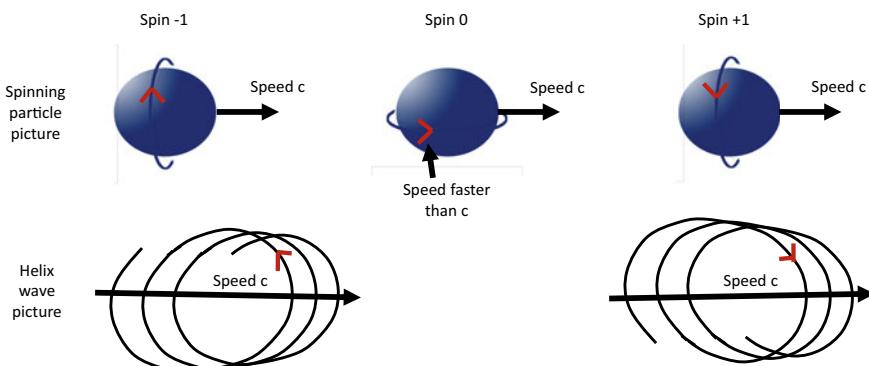


Fig. 9.9 Spin possibilities of the photon

But how can photons create the EM force? This would mean that a single electron would be spraying off photons all the time in order to produce the field, some of which are caught by other charged particles. This is the way a potential gets transferred to the other charged particle. But this would mean that the electron would continuously lose energy by spraying out photons all the time. This obviously does not happen. We will unravel this mystery later when we discuss virtual particles.

9.3 Path Integral

Before we can address the question of how the results of Sect. 9.2 create a force, we must pay some attention to the concept of the path integral. This will prove essential in understanding how a force comes about.

The basic concept for the path integral is the idea that a quantum consists of many waves, each of which gets from A to B via every possible route imaginable [Ref. 32]. Our understanding of a quantum really being made of waves makes this plausible: a wave extends in space in every possible direction. The probability of arriving at a point B is then determined by the interference of the wave at point B (see Fig. 9.10). So, the question becomes: how do we determine the probability of arriving at point B?

The path integral is a mathematical method to (1) figure out the phase of the wave at point B and (2) add up the phases of each wave that took a different route (see Fig. 9.10).

One problem in doing such a calculation is how to keep track of the phase when different routes are possible. In the path integral, this is done by having a clock move along with the wave. The clock times the phase of the wave. If we do not want to make a full calculation, but just look at a drawing to get a picture of the result, it is enough to show the wave and its phase.

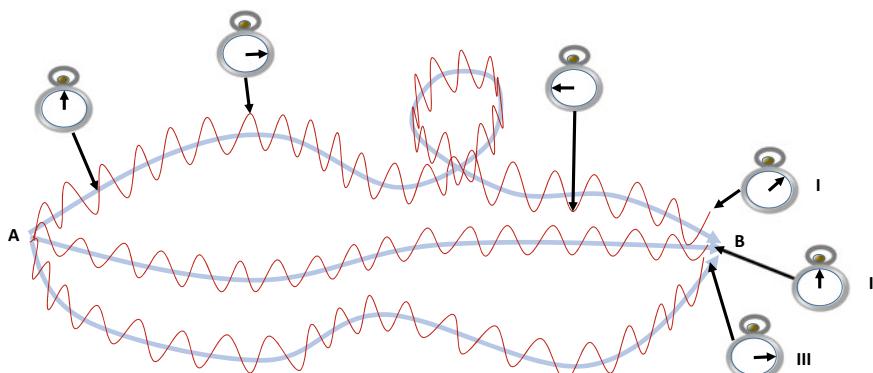


Fig. 9.10 Path integral: adding up the phases of all the waves that took different routes

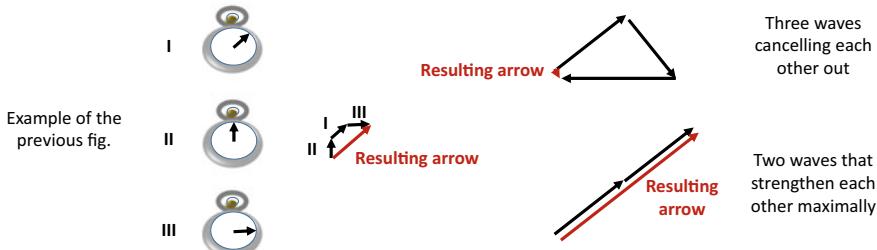


Fig. 9.11 Finding the result of the path integral at point B

In Fig. 9.10 three possible routes are shown. At some points along the top route the phase is indicated by the watch. Note that the watch has only one arm. The point is that the number of times the watch goes around does not matter. What matters is the phase the wave happens to be in at point B. So, at point B, each wave has a particular phase indicated by watches I, II, and III. The phases of these watches are added up to figure out how the waves interfere at point B. In this example they will show a net positive value when we add up the arrows of the watches.

The rules for addition are simple: you take all arrows of the watches of all waves ending at B (see Fig. 9.11). Then you place the back end of each arrow at the front end of the previous arrow. Having arranged all arrows this way, you draw the resulting arrow (the resultant) from the back end of the first arrow to the front end of the last arrow. The length of the resulting arrow indicates the extent of interference of the waves [Ref. 32]. Hence, it shows how large the field strength is at B. Suppose we are looking at just three waves and together they cancel each other out at point B. Then the resulting arrow will be 0: the waves cancel each other out. When two waves are “in sync”, their arrows point in the same direction and the resulting arrow has the maximal length: the waves maximally strengthen each other.

Now that we have an idea how to approach a path integral, let’s start by showing how a quantum that moves from left to right keeps moving in a straight line when there are no forces or other fields around (see Fig. 9.12).

Figure 9.12 shows that going from A to B in a straight line gives a positive interference between the waves. The resulting arrow is long. To show how this works we drew just a few paths. In reality there are many more paths and the paths further away from the straight line will show a larger phase difference. For instance, when the distance of a path differs by one wavelength from another path, the clock goes around a whole turn. In that case, it takes a wave one extra wavelength to take the longer path.

If we add up many more paths than shown in Fig. 9.12, we get the picture in Fig. 9.13. Here we see that the resulting arrow is still long. We also see that the length of the resulting arrow is mostly created by the green arrows in the middle. The green arrows are those that come from the paths that lie close to the straight line. Apparently, it is their contribution that makes the positive interference.

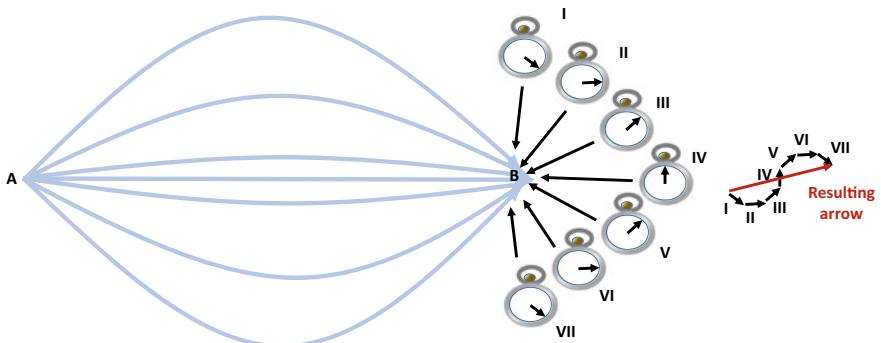


Fig. 9.12 Moving in a straight line

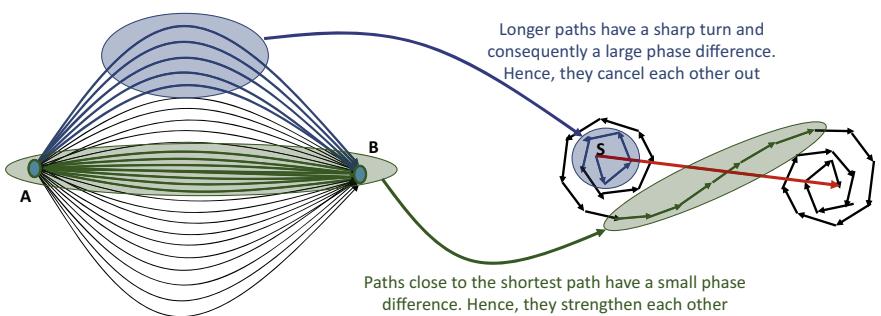


Fig. 9.13 Adding the phases of many paths from A to B [Ref. 32]

Now consider a more deviant path. The arrows for such a path and the paths nearby are shown in blue. It is clear that the resulting arrow for just these paths would be very small. In order to see that, just draw an arrow between the starting point S and the end point of any of the blue arrows. These are all short arrows, representing a small probability that such a path would be taken. This means that the waves going along that path will quickly interfere negatively. This is due to the fact that the phase difference between two paths gets larger when these paths lie further away from the shortest path. This you can see from the angle between two adjacent arrows on the right in Fig. 9.13: when two arrows make a sharper angle, it means that they have a larger phase difference. Remember the watches: a sharper angle between two arrows means that the watch for the wave on one path has moved on more compared to the watch for the wave on the other path.

Why would paths further away from the shortest path show a larger phase difference? The further away from the shortest path, the sharper the curve of the path. Compare this to walking on the road. When the road takes a sharp turn, it is much shorter to walk the inner side of the bend than the outer side. So, the more curved

the path, the larger the difference in path length between adjacent paths on the inside and outside.

You can also view it this way: for any path that is (much) longer than the shortest path you can find a shorter path that differs by half a wavelength and cancels that path out. Only the paths close to the shortest path have no shorter path that differs by half a wavelength that could cancel them out. This is reflected on the right in Fig. 9.13: the longer paths circle around each other when we add up their arrows. Only the shortest paths do not circle around and add up to a large resulting arrow.

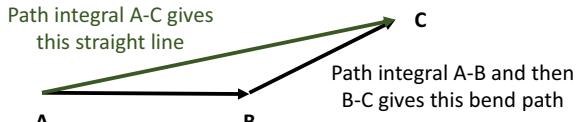
So, we conclude that only the waves close to the shortest path enhance each other. The waves taking longer paths will meet cancelling counterparts and are effectively excluded. Hence, when there are no forces, a wave will move along a straight line because that is the shortest path.

The shortest path is a straight line in this example. But you may argue that any straight line from A to B, or to C, or to D would show the same result! So why would a particle keep flying in the same direction? The next stretch may be in a different direction, as long as it is a straight line. For instance, take a particle that travels from A to B in a straight line and then to C in a different direction, also in a straight line (see Fig. 9.14, black arrows). Sure, but now consider this: you can repeat this process for any two points A and B! So, you may also apply the path integral to point A and C in one go (green arrow). Then it becomes clear that the path taken from A to C is different from the path taken first from A to B and then from B to C. This is inconsistent! The path integral must give the same result no matter what points A and B you choose. After all, it cannot be that a particle looks at you to see what points you choose and then decides what path to follow. The path a particle takes must be the same no matter which points you choose to calculate the path integral! Hence, the only consistent path that remains is a straight line. This must be the line following the original direction of the momentum given to the particle at the last interaction. So, you see that a sudden turn or angle in the path the particle takes is not possible when no forces act on it.

You can see this also in another way. Take the momentum density flowing in a particular direction. When there are no forces working on the waves, the flow of momentum goes in the same direction. Over short distances as well as over long distances the waves will keep interfering with each other. Hence, they will keep following the shortest path in the original direction of the momentum flow.

The path integral shows how a wave can take the shortest route. *The path integral explains why Fermat's principle works in the way shown in Fig. 5.4.* If you remember, I promised to clear up this mystery? It is the fact that the waves can take any route in combination with their mutual interference that forces them to take the shortest

Fig. 9.14 Path A-B-C with a sudden turn or angle is inconsistent with path A-C, and hence not possible



route. But why, you may argue, does the light follow a bent path in Fig. 5.4? We have just seen that, when there is no force, it should be going straight. Well, just wait and see!

The path integral also explains the straight line part of Newton's first law: "*In an inertial frame of reference, an object either remains at rest or continues to move at a constant velocity [hence, also in a straight line], unless acted upon by a force.*"

Since any object is just a bunch of waves that travel as a group, they too will move in a straight line because that is how the waves interfere with themselves. If we take the picture of objects as "particles" or "balls" moving through empty space, it is much harder to explain why they would move in a straight line. After all, how would they navigate in empty space? You can argue that they would go straight on when no force is exerted, but what constitutes "straight on" without a reference frame? Waves define "straight on" as the shortest route by means of their interference. "Particles" cannot do that unless they are made of waves.

9.4 How Does Symmetry Create a Force?

We saw in the previous sections that the coupling potential behaves like a spring attached to the wave. Let's now use the path integral to see how this gives us the appearance of a force.

We have seen that the coupling between the field of a charged particle and the electromagnetic gauge wave depends on the charge of the particle. Suppose we have two equally charged particles. When one of the particles shifts phase, it produces an electromagnetic wave that carries a potential difference. When that wave meets the other (equally charged) particle, the potential is being added. Consequently, the second particle gets an increased potential. It will experience that as an extra spring strength. That extra strength contains extra energy. Put differently, its mass increases.

When we add a spring (see Fig. 6.2), we increase the elasticity, leading to a lower speed and shorter wavelength (= higher momentum). However, we have also increased the energy. When two equally charged particles approach, they both lose momentum and gain potential, but the total energy in the system remains the same. Therefore, the effect of the gauge potential must be different than "just adding a spring". It must be adding a spring *at the cost of something else*. That "something else" can only come from the fact that their wave energy decreases. When the wave energy goes down, so does the wave frequency. This implies that their speed goes down and their wavelength goes up (lower momentum).

So, the effect of the potential is that the mass goes up due to the increased spring strength of the potential, but it does not lead to a shorter wavelength. The result is instead a longer wavelength, since the total energy in the system does not change. This shows that the effect of an exchange of gauge waves is not straightforward. In Chap. 10 we will discuss what such an exchange of gauge waves entails.

When the two particles are oppositely charged, the effect is opposite: when the electromagnetic wave meets the second (oppositely charged) particle, the potential

is being *subtracted* because the absorbing particle experiences a phase shift in the opposite direction around the circle. The result is that the particle's potential is lowered. It experiences that as a lower spring strength: its mass decreases. Moreover, the total energy remains the same in this system, so the consequence must be that their momentum goes up and their wavelength gets shortened.

Let's see what effect this process has on the waves and the path integral.

Two quanta with the same charge

The closer the two quanta are to each other the more gauge waves are exchanged. Hence, the more potential difference is exchanged. More potential equals a greater spring strength. A greater spring strength means a higher mass. So, the closer they get to each other, the higher their mass. The potential energy gets stored as extra mass!

From a momentum point of view, when the quanta enter each other's fields, their wavelengths get longer. The deeper they penetrate each other's field and the more gauge waves get exchanged, the more their wavelengths will get extended. This continues until the quanta come to a standstill and the wavelength is as long as it can get for such a quantum. So, they can only get as close as their initial kinetic energy (velocity) allows them to. When that is used up, they cannot get closer. Saying "used up" amounts to saying that their wavelength has reached its maximum length. The kinetic energy is then 0. At that point the quanta have built up the maximal amount of potential energy ("mass energy" or spring strength) they could with the momentum they had before. In Fig. 9.15 this means that the arrow of total energy is vertical along the axis of mass energy.

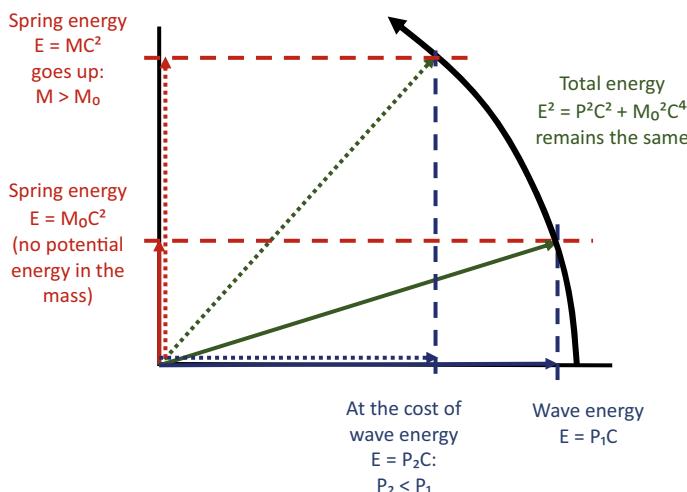


Fig. 9.15 When the total energy of the system stays the same, the increase in potential energy (mass spring) must be compensated by a decrease in wave energy (momentum)

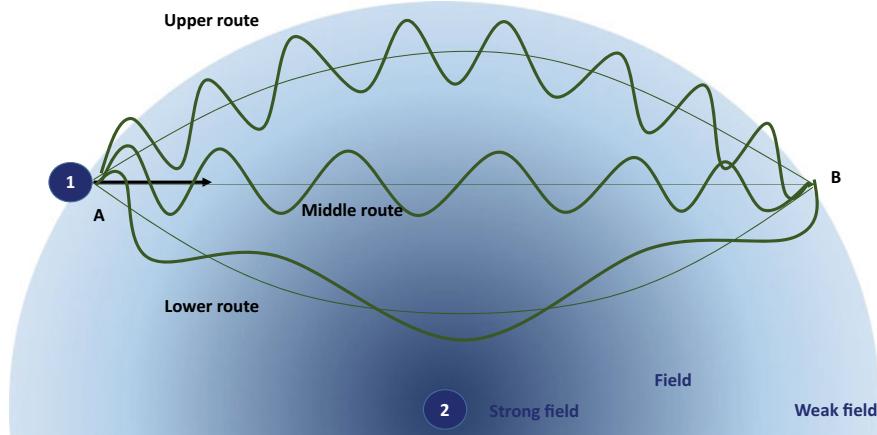


Fig. 9.16 Waves in the field of a quantum with the same charge: the interference between the waves leads to a shortest path that bends away from the other quantum, so we have a repulsive force

We end up with the fact that *quantum waves of equally charged quanta get a longer wavelength when they are close to each other, and a shorter wavelength when they are further away from each other.*

Now we let one quantum move past the other and see what the path would look like from point A to point B. For this we need the path integral (see Fig. 9.16). In the picture the waves that move close to the other quantum get longer and the waves that move further away get shorter. So, the shortest route gets influenced by changing these wavelengths. This picture is greatly exaggerated to give a better idea of what is going on. In the picture, the clock on the lower route makes fewer round trips than the clock on the upper route. So, in terms of numbers of wavelengths, the lower route is the shorter one. The lower route turns out to be a bend!

You may argue that you can find an even lower route that shows an even longer wavelength. Why is that not an even shorter path? The answer is that another effect must be taken into account. Remember that we discussed the impact of a curved route on the phase difference between adjacent paths? When the road takes a sharp turn, the inner side of the bend is much shorter than the outer side. Hence, the path along the outer side is significantly longer compared to the path along the inner side. So, this effect actually makes paths *longer* when they get more curved! The result is that, when we take a lower route in Fig. 9.16, we have two effects opposing each other:

- The wavelength gets longer, which shortens the path from a phase perspective
- The bend becomes more curved, which makes the path longer.

There is a path where the two effects are equally strong: this will be the shortest path. Any path lower will be too highly curved. Any path higher will have too short a wavelength. In any case we can conclude that the shortest path will be bent like the lower path in Fig. 9.16.

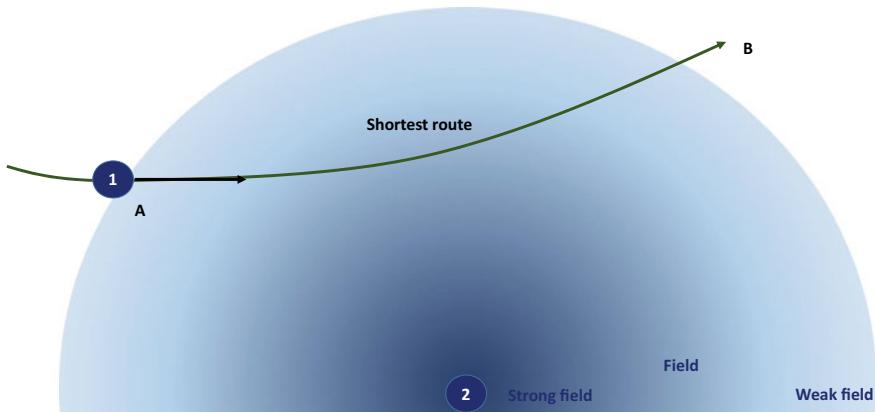


Fig. 9.17 A consistent, evenly curved path as would be described by path integrals taken between various points along the path

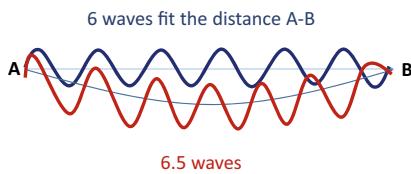
We thus end up with a curved path. The path bends away from the charge at the bottom. *Effectively, we have demonstrated a repulsive force!* Following this principle, why shouldn't the quantum start by moving straight to the right and then take a path down and up again? This seems strange, but here we see that we would have to look at path integrals between different sets of points to figure out what the actual path of the quantum is going to be. Here too, we need a consistent path where it does not matter which two points we take, because we end up with the same path.

Let's suppose that the initial position and velocity is as drawn in Fig. 9.17, with the quantum on the left and the velocity straight to the right. Different paths, starting before and after point A would all need to show the upward bending behaviour. Hence, we get a path that steadily bends upwards. This we recognize as a path that looks just as we would expect from a particle being deflected by the other charge, as described in the classical formulas.

Let's see what happens when the initial velocity is faster. The momentum we start with is higher. Hence, the wavelength is shorter from the beginning. So, what is the impact of a shorter or longer wavelength on the deflection process? Basically, the shorter the wavelength, the less difference in path length is required to get a specific phase difference (see Fig. 9.18). Put differently, when the initial wavelength is shorter, the path that differs by half a wavelength is much closer to the original path than when the initial wavelength is longer.

The consequence is that a wave with a shorter wavelength will be less bent than a wave with a longer wavelength. So, when a high velocity quantum gets deflected in a field, what happens to the two opposing effects of (1) making the wavelength longer in the field and (2) the curved path getting longer? When the quantum has a high velocity and thus a shorter wavelength, *it requires less bend in the path to compensate for the wavelength getting longer in the field*. In other words, the quantum

Short wavelength: to get half a wavelength difference requires only a slightly longer path.



Long wavelength: to get half a wavelength difference requires a much longer path.

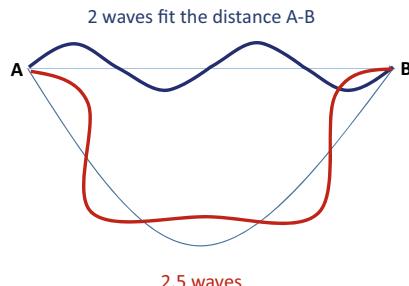


Fig. 9.18 The impact of a shorter or longer wavelength on the bending of the path

gets deflected with less bend. So, we see that, the higher the velocity of a quantum, the less it gets deflected.

It is also interesting to look at the impact of mass on the path. When we are looking at a larger object with a higher mass, the momentum is higher as well. In this case we also see that the wavelength gets shorter with a higher mass. Consequently, an object with higher mass gets deflected less: its path is less bent compared to an object with a lower mass. Here we see directly how mass is a measure of inertia: it is “harder” to change the direction of motion of a heavier object.

We were looking at inertia earlier in the book. Back then we said that *it was the (Higgs) springs that determined the inertia*. When the mass is greater, the springs are stronger and the frequency is higher (for the same velocity!). Hence, the wavelength must get shorter. Now we see how the path integral and the wavelength of the quantum confirm that the Higgs springs produce inertia.

Now suppose that the two quanta are initially close together and *not* moving. One of the quanta is fixed while the other is free to move. What happens? When there is no other quantum (situation 1 in Fig. 9.19), all waves going anywhere would interfere positively at exactly the same spot where the quantum was before. When there is another quantum (situation 2), the wavelengths away from the other quantum get relatively shorter, so the next spot the quantum will be in is not the same spot, but a little away from the other quantum. And this continues, with the first quantum each time further away from the other. In the process, the quantum picks up momentum. In this case too, the quantum experiences a repulsive force.

Two quanta with opposite charges

When the two particles were oppositely charged, we saw that the effect was opposite: when the electromagnetic wave meets the second (oppositely charged) particle, the potential is being subtracted. Therefore, the second particle gets a lowered potential when they get closer. It experiences that as a lower spring strength: its mass decreases. As we saw before, this implies that its momentum increases. Hence, its wavelength shortens.

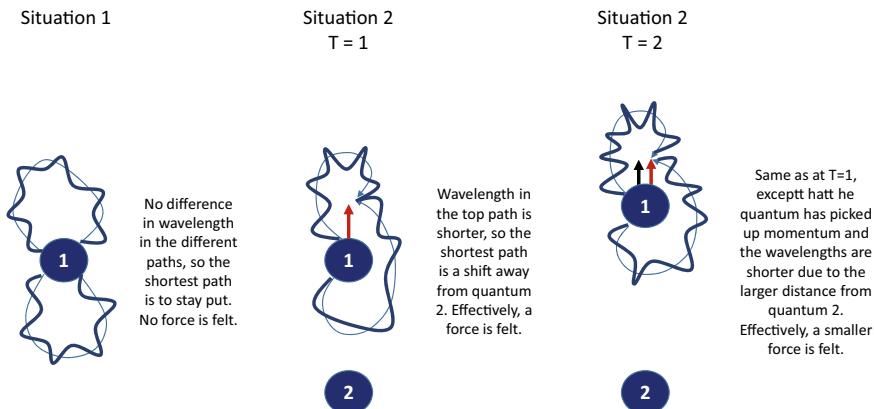


Fig. 9.19 A quantum in the vicinity of another equally charged and fixed quantum would start to move away from the latter. Situation 1 shows a single quantum. You can draw any path, but the shortest path is to stay put at the same spot. Situation 2 shows two equally charged quanta at different times. The shortest route is always to move faster and faster away from the fixed quantum

So, the wavelength of waves passing close by the other quantum will get shorter, while the wavelength of waves passing at a greater distance will get longer. Consequently, a quantum moving by another oppositely charged quantum will deviate from the straight path by bending towards the other quantum (see Figs. 9.20 and 9.21)

The effect of this process is an attractive force! If the quantum were to move straight towards the other quantum, its mass would get lower, its momentum higher, and its wavelength shorter. The quantum would be gaining velocity.

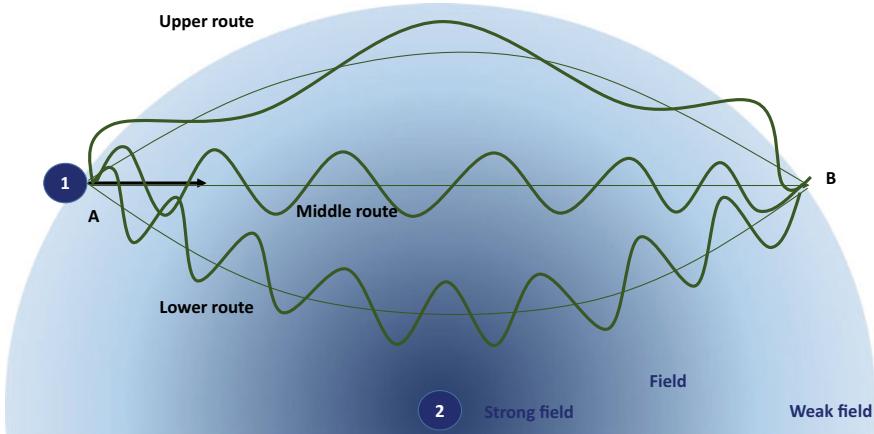


Fig. 9.20 Waves in the field of a quantum with opposite charge: this time the lower route has a shorter wavelength and the upper route has a longer wavelength. Consequently, the route from A to B will be taken rather than the upper route. This route bends as if quantum 2 attracts quantum 1

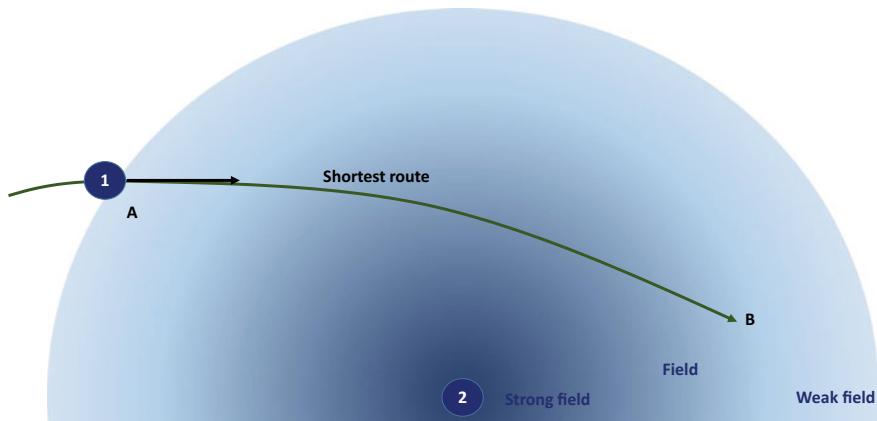


Fig. 9.21 Interference between the waves leads to a shortest path that bends towards quantum 2: we have an attractive force

9.5 A Constant Field and the Refractive Index

What happens in a constant field? In the examples we have seen so far, the field has a gradient: it changes in strength from the source outwards. This change in strength is responsible for the difference in wavelength of the waves that take different paths. When the field is constant, this difference is zero and the path does not get bent. There is no appearance of a force.

This is the case for the Higgs field. This field exists everywhere in space and is constant. So, Higgs does not bend the path of a quantum. It does act as a spring towards any field that interacts with it. Hence, Higgs gives these fields mass, but does not produce the appearance of a force.

Another example is the refraction of light in different media. As we saw before, the light gets a shorter wavelength in the medium since it interacts with the electrons in the glass. This changes the phase of the photon and causes the photon to slow down. The consequence is that the path gets bent. The extent of the bending is determined by the refractive index, defined as the ratio of the wavelengths in the two different media. When this ratio is high, the light will bend a lot at the interface between the two media, where the wavelength changes.

However, it only gets bent where the medium changes. Within a constant medium, light goes straight. In fact, it takes the shortest route between every pair of points you can specify along the way. The path integral has explained why (see Fig. 9.22).

So, we see that the path takes relatively more of the route with the long wavelength. Hence, it bends in the direction of the short wavelength, as we have seen in the previous examples. Inside the glass the path is straight, as it is also outside the glass.

A striking situation arises in the heat of the summer. If you are driving on a hot road you may see what looks like water on the road, but then turns out not to be.

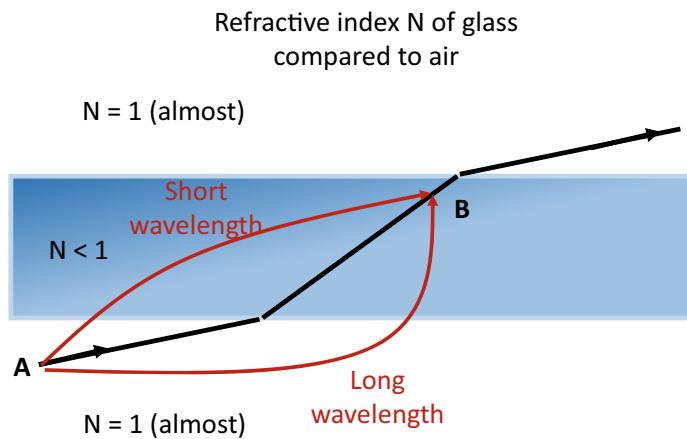


Fig. 9.22 Refraction of light according to the path integral

What you see in reality is a reflection of the sky in the road. Or better, the reflection of the sky in the hot air just above the road. How does that work?

The road gets heated by the sun and this in turn heats up the air just above it. The closer you get to the road, the hotter the air. Hot air has a lower density and therefore a lower refractive index. So, the air above the road can be seen as a medium whose density and temperature varies depending on the height above the road (see Fig. 9.23).

In cold air the light has a higher refractive index. It moves slightly slower than in hot air. Consequently, the wavelength of light is slightly shorter in cold air than in hot air.

This situation is comparable to a field that is not constant. And the result is that the light from the sky gets bent in the layers of air, much the same as a particle in an electromagnetic field. But when we talk about a particle in an electromagnetic field, we say that its path gets bent by the electromagnetic force. However, we do not

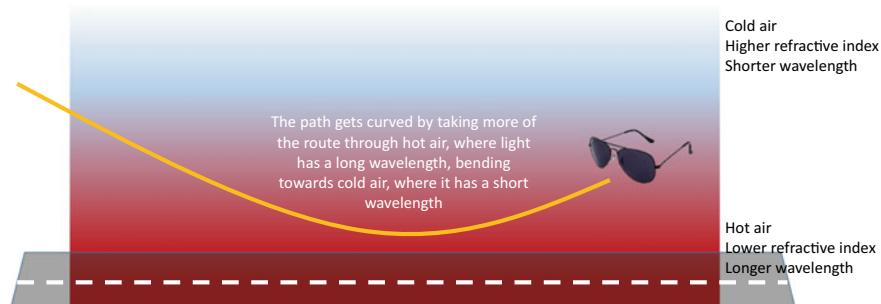


Fig. 9.23 Light getting bent by the varying air density above a hot road

say that about the air bending the path of light. While in fact these are very similar effects. Both are waves that go through a medium (field) that varies the wavelength of these waves so that they interfere positively along a curved path.

To me, this puts an end to the concept of a “force working over a distance”, and replaces it with waves moving through an inhomogeneous medium. So, the next time you feel a force (any force!), realize that this is just your waves going through an inhomogeneous medium instead of something actually pulling or pushing you!

9.6 Conclusion Regarding the Electromagnetic Force

The regular phase shift that a quantum may have has been identified as its charge. The phase shift produces a potential difference that propagates through the vacuum. The propagating phase shift has been identified as a photon. The potential difference that is propagated influences another quantum that feels this potential difference. The potential acts as a spring, increasing or decreasing the “mass”, or rather the “potential energy” of the other quantum.

These springs shorten the wavelength of the quantum or make the wavelength longer in different regions of the field. We have used the path integral to show that these waves interfere in such a way that the path of the quantum gets bent. This is the origin of the repulsive or attractive electromagnetic force that two charged quanta appear to exert on each other.

We have also shown that the bending of its path depends on the mass of the quantum. This is a direct consequence of the fact that the wavelength gets shorter when the total mass (total energy) increases. When the initial wavelength of the quantum itself gets shorter, its paths tend to bend less. Hence, we have seen the effects of inertia on how the path bends: inertia makes it “harder” to bend the path.

Other phenomena, such as the refraction of light in different media, can be explained in the same way. The path integral helps us explain how having a different wavelength in different media leads to a bending of the path at the interface between the media. Therefore, Fermat’s principle has been explained using the path integral. The path of light being bent in hot air shows exactly the same behaviour as a quantum in a varying field with the same type of curved paths. For quanta we are used to thinking of such a bent path as a “force”, but now we can replace that concept by waves going through an inhomogeneous medium!

Chapter 10

Propagators and Virtual Particles



In the previous section we showed how a phase shift produces a potential difference that is propagated away and can interact with another wave and hence produce a force. This also means that the phase shift causes a wave that transports energy and momentum away. So, when a wave causes phase shifts all the time, it sends off energy all the time. This cannot be possible, since the wave would be losing energy all the time and would soon cease to exist.

So how does this work? Let's take the example of an electron. Basically, the electron that produces the phase shift turns into a virtual electron and a virtual photon: it splits the phase change off. Some of the energy of the electron will go into the photon that propagates the phase shift away.

Then one of two things happen: the virtual photon either gets recombined with the electron, thereby restoring the phase shift, or the virtual photon gets absorbed elsewhere and the virtual electron lives on with lower energy. So, the mystery of losing energy all the time seems solved by saying that each disturbance that does not get absorbed by another quantum gets absorbed back by the quantum that emitted it.

However, this triggers some new questions:

1. The electron does not always have enough energy to produce a real quantum, but the phase shift happens nonetheless. How can a proper quantum (the electron) split into two virtual particles? What are these virtual particles, if they are not proper quanta? How can they exist when we saw before that only quanta of a particular energy can be true excitations of the field? After all, when a particle gets absorbed, only a little energy loss is enough to make the wave function collapse. Then, why would that not be the case when splitting off a photon?
2. When the photon gets absorbed elsewhere, how can the electron live on in its virtual state of being?

Let's start with the first question by looking more closely at what is actually happening during a phase shift and what it really means to be a real particle or a virtual particle. I will answer the second question a little later when we discuss the characteristics of virtual particles on the mass shell.

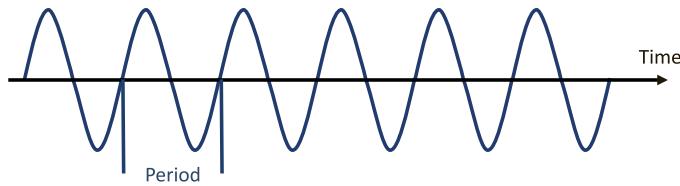


Fig. 10.1 A wave with one frequency is a single wave on the time axis. The frequency is equal to 1/period

Let's begin by looking at the wave when a phase shift takes place: it is no longer a perfect wave (see Figs. 9.5 or 9.6)! There is an ugly step in the original wave, so it can no longer be described by one wave with one momentum. It must change momentum. The momentum goes into the newly created wave. The sum of the momenta and energy remains the same.

But the wave started in the gauge field does not need to meet the conditions to be an actual photon, i.e., the energy needed to produce a photon at that frequency does not have to be available. So, the relation between energy and momentum does not need to apply to this wave. However, the consequence is that this wave (since it is not a proper quantum) cannot last. Why would that be?

When we excite a harmonic oscillator, we can do that by putting in exactly the energy hF (h = Planck's constant). This represents one field quantum. Such a quantum is a real quantum with exactly the (mass) frequency F in the corresponding (matter) field. This exact frequency implies that the quantum will continue forever on the time axis (see Fig. 10.1).

When we discussed the uncertainty relations we saw that, when the energy (frequency) is known exactly, we cannot establish any limited lifetime. Put the other way around, if the lifetime is limited, the energy (frequency) cannot be known exactly. When we are talking about a field quantum, we can know exactly what frequency it has, and therefore, it must be living a long life. This is what we call a real quantum, or a real particle.

A real particle is like a resonance on the tuning fork. A resonance is its natural motion. Like a child on a swing, it will swing with different amplitudes, but always the same frequency. Once excited, a tuning fork will keep producing its natural tone for a long time. Eventually, it will dampen down since it loses energy by hitting air all the time while giving off its tone. When a quantum produces phase shifts all the time, something similar happens. But there is one big difference: the tuning fork has plenty of possibilities to pass on its energy, but a quantum has not. It cannot produce a real photon since the phase shift generally does not provide enough energy to do so (or the right combination of momentum and energy as we will see). The only thing it *can do* is disturb the gauge field. So, it produces an “incomplete gauge photon”, or better, a disturbance that cannot sustain itself. This means that the disturbance has to die off somehow.

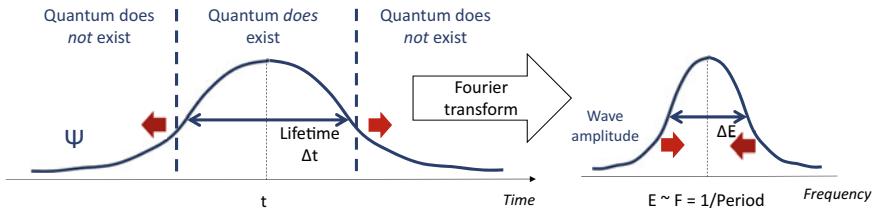


Fig. 10.2 A short lifetime leads to a wide variety of frequencies in the wave packet, hence a wide variety of energies. When the lifetime gets longer, the variety of frequencies shrinks (red arrows) and the energy of the wave packet becomes sharper

So, the gauge quantum has a limited lifetime. It exists only for a limited period of time. On the time axis this looks like Fig. 10.2.

We can only produce a shape like this when we sum up waves of different frequencies. The shorter the lifetime, the more frequencies are required. This is how an incomplete quantum, a disturbance, can exist for a limited time, as a wave packet with unclear frequency.

We can compare that to an attempt to make the tuning fork ring with a different frequency. That requires a lot of work. Or put a child on a swing and try to make the swing go at a different frequency than its resonance, or natural frequency. You would have to put in a lot of effort to do so. You would have to push the swing and catch it at some point, then push it again. In order to make it go at a different frequency, you would continuously need to interact with it. Upon each return, you would need to stop it and push it again. In fact, you can only force a harmonic oscillator into a non-natural motion or a non-natural frequency by disturbing it, and such a disturbance is essentially different from its natural motion.

Something similar happens when phase shifts take place. The phase shift is actually a disturbance in the electromagnetic field. This disturbance cannot be compared with a quantum. A quantum in the electromagnetic field (a photon) contains an energy in line with its frequency. But the phase shift need not produce that combination of energy, wavelength, and frequency. Therefore, the disturbance is not a quantum and it cannot consist of a single wave. Such a disturbance is not stable. The waves in the wave packet will eventually cancel each other out.

But the disturbance does carry energy, so it can only die out by being reabsorbed, i.e., before the oscillation is over, it is caught again by the electron that produced the phase shift in the first place. During this time, both the electron that is the source of the phase shift and the gauge photon have a different energy and momentum. However, when we sum these up, their total energy and momentum is equal to that of the original electron before the phase shift happened. So, the total energy and momentum does not change. The electron and the photon remain linked, so that the wave function does not collapse. Before that could happen the photon and the electron are recombined.

An interesting thing happens when the gauge photon is caught by another quantum. Then that other quantum absorbs its energy and momentum, and the original electron

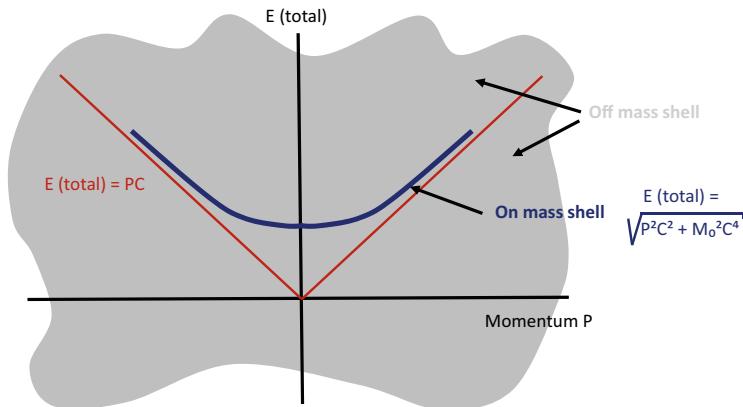


Fig. 10.3 Disturbances are off mass shell and quanta are on mass shell. Quanta follow a strict relativistic relation between energy (frequency), momentum (wavelength), and mass, namely, the dispersion relation. Disturbances do not

has to go on without it. So effectively, a transfer of energy and momentum takes place. In order to see what happens to the electron, we will have to take a closer look at our old friend, the mass shell.

On or off the Mass Shell

Since the lifetime of a disturbance is short and its frequency is not defined properly, its energy cannot be defined properly either, and the disturbance must be made of a bunch of waves. The momentum to go with each wave in the wave packet is related to the frequency by means of the velocity of the wave, but this may be any velocity. Then there is the velocity of the wave packet, which can be any velocity when the frequencies, velocities, and wavelengths of the individual waves can vary so much. Some waves can be so far off resonance with the springs that the springs have no time to react, resulting in a very low *effective* mass.

If we look at the mass shell (see Fig. 10.3), a quantum on the mass shell has a strictly related mass, momentum, and energy. The relation between them is mathematically fixed in a relativistic dispersion relation. This does not work for disturbances. They may have any relation whatsoever (the grey area in the picture). If you look at Figs. 9.5 or 9.6, you see that when a quantum shifts phase, both the quantum wave and the gauge wave do not really look like waves. The quantum wave shows that ugly step and the gauge wave is more like a step travelling away than a nice sine wave. So, when a quantum shifts phase, its wave goes off mass shell and the gauge wave is off mass shell too. Only the characteristics of quantum waves + gauge waves together remain on mass shell, as their sum still represents the original quantum.

In the previous chapter we discussed how an exchange of gauge waves could increase the potential (mass energy) of an absorbing particle. When we increase the spring strength, we expect (in a perfect wave) its wavelength to shorten (see Fig. 6.2). We discussed the fact that the opposite happens (the wavelength is extended) because

the total energy in the system could not change. Now we see that such a thing can happen because the total energy of the wave is split over two imperfect waves, two disturbances that are off mass shell. For such particles the usual perfect wave relations do not work. And so it can happen that such a splitting off and absorption could increase the spring strength while extending the wavelength.

After some time, the quantum waves start to look like a proper wave again (when the step of the phase shift becomes history), so when the gauge waves are absorbed by another particle, the original emitting quantum becomes a real quantum again, but now with different momentum and energy. When the quantum waves reabsorb the gauge waves the quantum emitted before, you can imagine that they will fit well to recombine into the original wave. Let's say that the original step is cancelled by receiving that step back in exactly the same way.

When the gauge wave is absorbed by another quantum and a transfer of momentum takes place, the momentum of both quanta will have been uncertain for a short while. This means that during this period of momentum transfer we need to consider the total waves of each quantum (including its disturbances) as changing waves with a changing wavelength. How can we look at a changing wave? Just as with any imperfect sine wave, we will need to build such a change from a whole bunch of sine waves. A bunch of sine waves with different wavelengths creates a lack of clarity regarding the momentum of the quantum during this transition. At the same time, it creates more clarity about the positions of the two quanta! Compare this with Fig. 4.4. Even if the bunch of sine waves do not create an exact spike, they will sum up to a total wave with a different probability in different positions in space. So basically, any interaction produces information about the position of each quantum. This is an important mechanism for understanding the “collapse of the wave function”. We will get back to this in Sect. 12.1.

A disturbance of a field does not have the same mass as an excitation of the field (an actual quantum). As we can conclude from the grey area in Fig. 10.3, it can have any mass, including a negative mass-squared, sometimes called an imaginary mass. In this sort of situation there is no natural definition of the mass of a particle [Ref. 30, p. 463]. I cannot presume to offer you any image of such a wave. However, if this can exist, it seems plausible that we could even create disturbances of the electron field that are massless. We will use this idea later.

10.1 Time Order of Events and Feynman Diagrams

We have been discussing quanta that can produce disturbances which can either be reabsorbed or caught by another quantum. In the latter case momentum is transferred, the paths change, and we experience a force.

As we will see, there can be many such processes. In calculations, the multitude of possible disturbances and interactions lead to long and complicated integrals. In order to keep track of the disturbances and interactions, they are worked out using simple

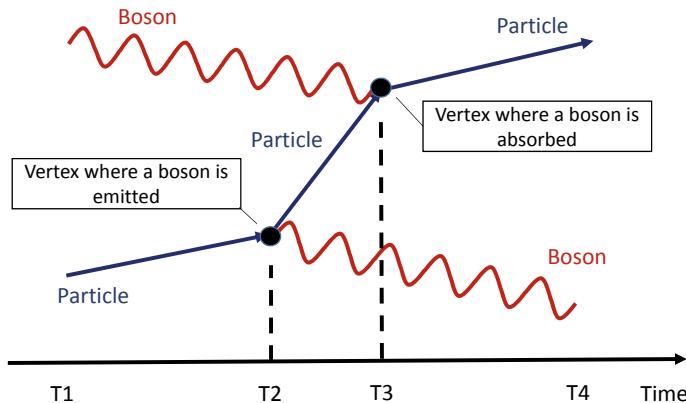


Fig. 10.4 Example of an interaction: a particle changes path by sending and receiving bosons

diagrams. These diagrams are called Feynman diagrams, after the great twentieth century physicist Richard Feynman, who developed this approach [Ref. 2, 32, 56].

A Feynman diagram generally consists of several components:

- Real quanta entering the diagram, depicted as incoming external lines
- Interaction vertices, depicted as points where three or more lines interact
- Disturbances, depicted as internal lines between two vertices
- Real quanta leaving the diagram, depicted as outgoing external lines.

Bosons are depicted by wavy lines. Fermions are depicted by straight lines.

Let's take a look at a first example (see Fig. 10.4). A real particle comes in, emits a real boson, moves on, absorbs another real boson and leaves.

In this diagram we also show a timeline so you can see in what order the interactions took place. The immediate question that pops up is: what would another observer see if they were moving at a different velocity? Relativity shows that the time-order of events as perceived by one observer may be quite different for another observer. So, the other observer may see T2 and T3 reversed! Let's see what that diagram looks like (Fig. 10.5).

This diagram is essentially different! If we follow the timeline, the particle and the boson first enter the scene. The next thing we see is that the boson splits up into a particle-anti-particle pair at T3! That is certainly not what the first observer saw in Fig. 10.4. The particle and anti-particle move on until the anti-particle meets the particle that originally entered the scene at T2. When they meet, they annihilate each other and produce a boson that leaves the scene. The first particle that was created as part of the pair also leaves the scene.

But what is an anti-particle? Essentially it is the opposite of a particle. For instance, the anti-particle belonging to an electron is called a positron. All particles have an anti-particle. Generally, the only way to create a particle is by creating both the particle and its anti-particle. This is called “pair production”, since a pair of particles

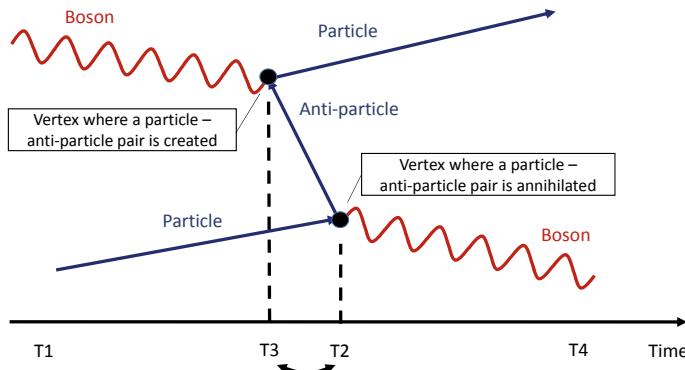


Fig. 10.5 The same interaction as perceived by a different observer: T_2 and T_3 are reversed

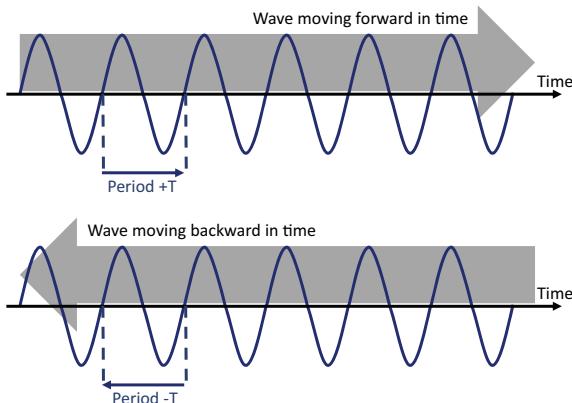
is created. A particle and its anti-particle can also annihilate each other when they meet. At that point, the energy is used to create another particle such as a high energy photon that will carry all the energy away.

So, what the first observer saw as one particle having some interactions with bosons, the second observer saw as a complicated pair production and annihilation. And all this because the time order of two events was reversed [Ref. 35, p. 120]! This problem can be resolved if we assume that, in Fig. 10.5, *the anti-particle is in fact the particle travelling back in time!* If we interpret the situation like this, the two observers can at least agree that they are both talking about one and the same particle. It's just that the second observer sees that particle move back in time. Pair production is in that case not really "production", but just a particle reversing its time direction. The explanation of anti-particles as travelling back in time is called the Feynman–Stueckelberg interpretation [Ref. 30, p. 145].

One conclusion that can be drawn is that, apparently, waves do not need to be restricted to going forward in time. Hence, when a particle is represented by a wave, the related anti-particle is represented by a wave that travels back in time. An anti-particle wave must then have a negative frequency (see Fig. 10.6). Looking at the time axis, the time required for one wavelength to pass is positive for a normal particle, but for its anti-variant, time goes the other way, so the time required for one wavelength to pass is negative: it is $-T$. The frequency is then $F = 1/(-T)$ and is also negative.

An anti-particle also has an opposite charge. So, when a particle has a negative charge, its anti-variant has a positive charge. Why an opposite charge? Charge was identified as the ability to shift phase and produce (or absorb) a gauge wave as a consequence of that shift. When we picture this process (see Fig. 10.7), we see that, when a disturbance approaches a normal particle wave, the potential of the gauge wave gets transferred over a short period of time when the disturbance wave passes by. But when it approaches an anti-particle wave, the clock of that wave goes backwards. The effect is that the disturbance wave actually moves in the other direction as seen by the anti-particle wave! Just like playing a film backwards. When it does so, the disturbance wave has exactly the opposite effect. So, when the disturbance would

Fig. 10.6 The negative period of a wave moving backward in time leads to a negative frequency



decrease the mass (potential energy) of the particle as well as shorten its wavelength, it would always do exactly the opposite to the anti-particle. Then we can also expect the change in direction of the path to be exactly opposite. So, where a particle would be attracted, the anti-particle path would bend the other way as though it were repelled. Therefore, in the most general sense, anti-particles always have exactly the opposite charge compared to their normal versions.

Another way of seeing this is to consider that the opposite effect of the disturbance is an opposite phase change. And from Fig. 9.8, we see that an opposite phase change is the same as an opposite charge.

Note, however, that an anti-particle has a positive energy in our frame of reference despite its negative frequency! You can see this by annihilating an electron against a positron. The energy produced is equal to that of two electrons. So, in our frame of reference, we meet particles and anti-particles that all have a positive energy, even though the anti-particles have a negative frequency in our frame of reference.

So, we have witnessed some compelling arguments for anti-particles to be just particles that travel back in time. But is this true? Opinions differ, but to date there has been no experimental verification that anti-particles would (not) be travelling back in time. Without such verification either way we cannot be sure. Personally, I find the arguments convincing enough to seriously consider this as a hypothesis. Moreover, there is a symmetry problem. If we are to treat time as a symmetric dimension, any excitation of a field would lead to a wave in both time directions. So, the creation of a particle would lead to a wave in the positive time direction as well as a wave in the negative time direction. Compare this to throwing a stone in the water: the wave will go equally in all directions. Symmetry means that no particular direction is singled out for the wave to go. Taking this to one space dimension such as a rope, when we pull up the rope in the centre, a wave is started in both directions. Just try to imagine a situation where the wave only goes to the right along the rope, but nothing happens on the left. As long as the situation is symmetric we will always have a wave going in both directions.

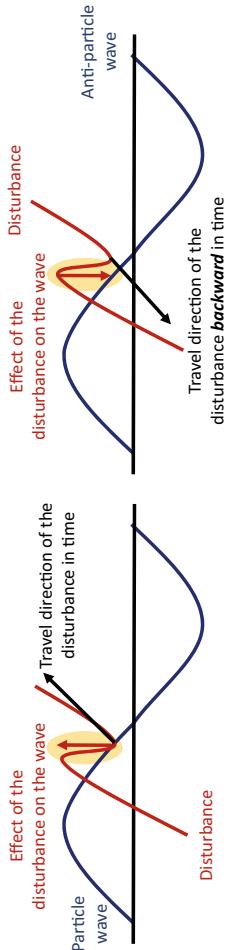


Fig. 10.7 Particles and anti-particles experience the disturbance wave in opposite ways

Consequently, when the excitation of a field leads only to particles and anti-particles in the positive time direction and nothing happens in the negative time direction, *time cannot be symmetric!* Hence, if anti-particles did *not* travel back in time, time would not be symmetric. That means that there would be a fundamental asymmetry in the universe. If that were the case, I would have to assume that the symmetry in space–time was broken around the beginning of the big bang. In Sect. 18.5 we will see that the weak force shows some asymmetric behaviour that impacts time symmetry as well. This is indeed a consequence of symmetry breaking in the early universe. However, the extent of this symmetry breaking so far seems very limited. It falls far short of explaining how a field excitation could lead to a wave in only one direction of time. So, to date there is no evidence that the symmetry of our time dimension is broken so completely that it causes the arrow of time to go in only one direction. Personally, I prefer the assumption that time is generally symmetric (with some exceptions that are corrected for in CPT symmetry, see Sect. 18.5), and consequently that anti-particles actually do travel back in time.

What does this mean for the way we experience the world around us? Since we perceive the universe as moving in one time direction (e.g., entropy increases, an egg breaks into pieces and does not come together again), we need to explain how that is possible in a time-symmetric universe. You may also wonder what effect it would have if anti-particles really did travel back in time. Would that influence causality? I will go into these questions in a separate section on the “arrow of time”.

So now we end up with Feynman diagrams that are different depending on the observer. That is not good. We need a theory that does not depend on the observer! Such a theory is said to be Lorentz invariant. That means that the theory does not change under a Lorentz transformation. A Lorentz transformation is a change from one frame of reference to another, or put simply, it is a change to a different observer moving at another velocity. After all, physics should not change when we move at a different velocity. That is what we learned from relativity.

So how can we make Feynman diagrams Lorentz invariant? What we have to do is take the time order of events out of the diagram. We can only do that by including both possible processes (of Figs. 10.4 and 10.5) in the calculation. So, the diagram is simplified by taking the time axis out and by including all possible ways the interaction can take place in the calculation.

This means that a meaningful calculation of an interaction can only be obtained by adding all possible ways that the interaction can take place. Different observers will see different ways, but they have to include the ways of all other observers in their calculation. Each observer will then end up with the same conclusion. Doing this makes the diagram Lorentz symmetric, which makes it easier to describe interactions properly. If we are combining interactions, we do not need to worry about whether we have taken all possible variations into account. They are already added up in the Feynman diagram. This “adding up all variations” is done in what is called the propagator of the interaction (see next section).

It looks like building a house from symmetric (Lorentz symmetric!) bricks rather than from asymmetric bricks [Ref. 35, p. 126]. A lot easier! When calculating a type of interaction, we do not have to do a whole bunch of math that has already been done

before. We simply draw the Feynman diagrams that apply to the interaction, apply the Feynman rules to the lines and vertices (each Feynman rule adds a multiplication factor), and end up with the formula that describes the interaction.

Energy and momentum are conserved at any vertex in the Feynman diagram. So, the change in momentum of the incoming particle towards the outgoing particle has to equal the momentum carried away by the disturbance.

10.2 Propagator

Let's look at virtual photons. We discussed before how the phase shift produced by a source (a particle) creates a potential difference that propagates through space. The word "propagates" has a special meaning, as the solution that describes the potential difference "wave" is also called a propagator. *It describes the propagation of a packet of waves that summed up can have a different character than an actual quantum.* It can be off mass shell and it can have a limited lifetime. Basically, a propagator describes how a *disturbance* moves through space–time.

There is something funny about the propagator, as it can describe a real quantum as well. The real quantum is in fact a special case of a disturbance. A real quantum is the "pole" in the propagator. This is a special condition in the propagator with special characteristics. In this case the solution becomes a single wave, on mass shell, with unlimited lifetime. The quantum satisfies conditions that do not apply to the rest of the propagator, but only at the pole. Being the pole in the propagator, a real quantum is also called a *resonance*. A real quantum is therefore a special case of a disturbance: the true excitation of the field. When a disturbance looks more like a real quantum, the propagator describes a higher probability of finding such a disturbance and a longer lifetime. However, as we have seen before, a real quantum never really exists. It will always interact with other fields and produce disturbances. When it does, it is no longer the ideal real quantum.

When it comes to propagators, there are different types. There is the free propagator, which describes just a virtual particle (disturbance) going from A to B without interaction. There is also the full propagator, which describes the total propagating process of an interaction and all virtual particles that take part in it. Interactions can differ, e.g., by taking into account the number of particles entering the process and the number of particles leaving. There is a full propagator for each possible number of particles involved in an interaction.

For instance, when one particle enters and one particle goes out, the full propagator describes a process with two particles and is called a 2-point propagator. This can be used to describe a disturbance that is created by the incoming particle and gets absorbed back by it. A 4-point propagator is a full propagator that describes an interaction with two particles entering and two particles leaving ($2 + 2 = 4$). This can be used to describe the interaction processes between two particles.

The full propagator in space–time is defined (conceptually) by:

propagator = sum of waves that each have a different momentum \times amplitude of each wave

The amplitude refers to the “probability amplitude” and not to the “height” of the wave. It represents the probability of finding a wave with this particular momentum in the wave packet. Hence, the (probability) amplitude of each wave depends on its momentum. In the case of the 2-point propagator, the amplitude for the disturbance waves (the virtual particles) also depends on the following factors:

- $\Sigma(p)$: tells you how much each wave with momentum p in the packet interacts with other fields in the vacuum.
- $\Gamma(p)$: tells you the decay rate of each wave with momentum p in the packet. The decay rate limits the lifetime of the wave at momentum p .

Let's take the two-point process: a particle enters, it sends off a disturbance which it reabsorbs, and the particle leaves. The probability for this process to happen is roughly found by multiplying the incoming wave by the 2-point propagator and the exiting wave:

$$\text{probability(2-point process)} = \text{incoming wave} \times \text{propagator} \times \text{outgoing wave}$$

This process can be represented in a diagram (see Fig. 10.8) which shows the creation and re-absorption of the disturbance. The diagram also shows how the momentum stays the same at each stage of the process. The same goes for the (average) energy.

When a particular wave with momentum p does not interact with other fields and has an infinite lifetime, both Σ and Γ are 0. This is the condition for the probability of such a wave to be 100%. This is the situation at the pole, for a real particle: the wave is perfect, not interacting with other fields, not creating any disturbances, and with an endless lifetime. When Σ or Γ are not 0, the wave is not perfect, in which case it interacts or creates disturbances and it does not have an endless lifetime: it is a disturbance.

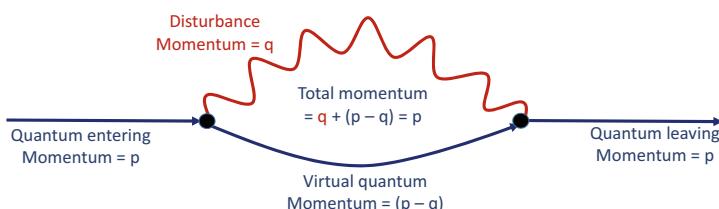


Fig. 10.8 Diagram for the creation of a disturbance and its re-absorption. As you can see, at every stage the momentum of the whole thing is equal to p , the momentum at the beginning

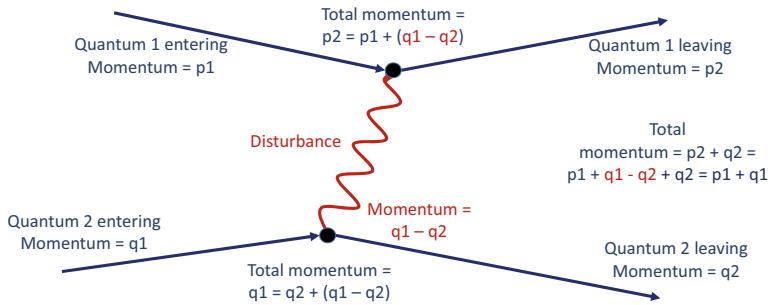


Fig. 10.9 Diagram of a 4-point process: interaction between two particles. The disturbance takes momentum from q_1 away and adds it to p_1 . Hence, the momentum is conserved at each vertex in the diagram and as a whole

In Fig. 10.8 the quantum sheds some of its energy and momentum in the disturbance. We call this energy *self-energy* when the disturbance is re-absorbed by the particle. This process is described by the discussed 2-point propagator. Self-interaction is a special process that can cause infinities to appear in the mass of the electron. We will address this mystery in the section on renormalization, where we will explore this type of interaction further.

Now let's look at the propagator for *two* particles entering and *two* leaving (the 4-point propagator). This describes an interaction. The probability for this process to happen is thus

$$\text{probability} = \text{entering(wave 1} \times \text{wave 2)} \\ \times 4\text{-point propagator} \times \text{exiting(wave 2} \times \text{wave 1)}$$

This process is described in another diagram (see Fig. 10.9).

Up to now we have described the two possibilities that can happen to a disturbance created by a particle:

1. It can be re-absorbed, which is a self-energy process. The disturbance is described by a 2-point propagator.
2. It can be absorbed by another particle, which is an interaction process. The disturbance is described by a 4-point propagator.

In general, we find that the propagator describes a whole bunch of waves that can start and end and need to be summed up. So, this bunch of waves must interfere with each other and together form a disturbance in the field. The only certainty baked into this is that the energy and momentum density of the sum of all waves entering equals the energy and momentum density of the sum of all waves leaving.

In order to describe *all* waves that make up the interaction, the propagator must add up all possible ways an interaction can develop. All possibilities count and add up to the total probability for the interaction described. So what possibilities are there and how can we see that the propagator accounts for all of them? For instance, the

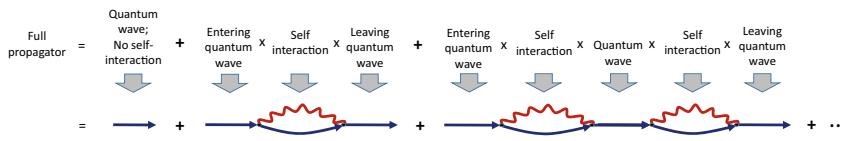


Fig. 10.10 The full propagator can be written as a series that can be translated directly into diagrams

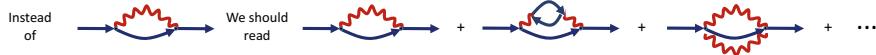


Fig. 10.11 All possible ways one self-interaction could develop

2-point propagator that describes the self-energy of a particle can be written as a series (see Fig. 10.10).

As you can see, the full 2-point propagator takes into account the fact that there may be any number of self-interactions when a particle moves from A to B, including no self-interactions. The diagrams in Fig. 10.10 are not entirely correct. Every single self-interaction actually represents a series itself. It represents a series of possible sub-interactions, such as those discussed with regard to the time-ordering of events (see Fig. 10.11).

In Fig. 10.11 we find for instance the possibility that a disturbance (the red wavy line) may in turn disturb another field and create a virtual particle and anti-particle pair before continuing and being reabsorbed by the initial quantum. There is also the possibility that two disturbances occur at the same time, and we can go on like this forever. The probability of each variation falls quickly when it contains more nested disturbances.

The series in Fig. 10.10 adds up to the full propagator. It shows the number of times the interaction can take place. Each interaction can be associated with a series of diagrams as in Fig. 10.11, showing all the possible ways the interaction may evolve. So now we see that the n-point propagator describes all possible ways that an n-point interaction can take place in one formula. It can be represented by a combination of diagrams that each describes a possible version of the same type of interaction. Summarizing:

$$\text{n-point propagator} = \text{sum of all possible Feynman diagrams with } n \text{ external lines}$$

Although all Feynman diagrams need to be taken into account, not all contribute to the process by the same amount. How much a diagram contributes to the process depends on how far off shell the disturbances are that produce the force. The further off shell, the lower the probability that the process represented by that diagram will happen. Hence, the lower the contribution of that diagram to the total process.

So, when is a disturbance further off shell? Let's look again at Fig. 10.9. This is a high probability contribution as the disturbance could very well be (close to) on shell.

For example, it may be a photon that transfers the momentum and energy cleanly in such a way that the energy transferred is equal to the momentum transferred times c ($E = PC$). But the versions we see in Fig. 10.11 may be less probable. Take for instance the situation where the photon splits into a particle–anti-particle pair for a while midway. These two particles must have an energy and momentum that adds up to $E = PC$ (valid for the photon when it is close to on shell). But if they were on shell, there would have to be a different relation between energy and momentum for these particles! So, they *cannot* be on shell and are in fact far off shell. Hence, the contribution of this situation to the total process is much smaller.

To avoid confusion, the propagator we have been talking about is referred to as the Feynman propagator. In general, the Feynman propagator describes the exchange of virtual particles. Or more precisely, it describes a packet of waves going from the interaction vertex A to interaction vertex B. Often, it is just called “the” propagator. However, propagators also exist for real particles that have a mathematically very different form. They describe the probability for a real particle to get from position A to position B. These types of propagators bear no relation to the propagator we have been talking about in this section. They are used in the context of quantum mechanics, while the Feynman propagator is used in quantum field theory.

Suppose we put a source (a charged particle) in the vacuum. When we do so, the electromagnetic field gets disturbed and becomes inhomogeneous. The consequence is that other charged particles moving through the disturbed field will experience a force and we can use that force to move something or to create induction (i.e., an electric current). Examples are:

Transformer: the electromagnetic field produced by one coil (source) induces a current in another coil. This is what happens in your toothbrush loader when you have a separate loader (source) that goes into the power socket and a toothbrush to place on it. The disturbances created by the loader induce a current in a coil of the toothbrush that loads the battery inside the toothbrush.

Near-field communication: the source could for instance be a payment machine that produces an electromagnetic signal that can be picked up by your bankcard. The bankcard has a chip and an induction coil. The card gets both energy and the signal from the payment machine and uses this to send an answer back to the machine. All this sending is done using disturbances.

MRI (Magnetic Resonance Imaging) scanner: the scanner is the source here, producing strong disturbances. These disturbances are picked up by certain molecules in your body (e.g., water molecules). The way they are picked up is called a resonance (but not in the same way as we discussed in this section), hence the name of the machine. The molecules get excited (in particular, their spin direction changes). They respond to this by falling back to their original state after some time. When they do so, they produce (real!) photons that are measured. The point is that the field created by the machine has a gradient, so only at a particular distance do the molecules pick up exactly the right energy from the field to get excited. By changing the field strength, that distance changes. So, the machine produces a field, measures a slice, increases the field, and measures the next slice. In the end, the machine can

indicate where the given molecules are located in your body in smaller or greater concentrations.

10.3 Relation Between Virtual and Real Particles

One of the differences between a real and a virtual quantum is in its lifetime. So, how about a real photon that gets absorbed quickly? It would have had a very short lifetime so this photon too should consist of a number of waves and frequencies and its energy would be uncertain, right? This may look like a disturbance. The main difference, however, is that the disturbance would still satisfy a different relation between its mass, energy, and momentum. For instance, the disturbance exchanged between an electron and a nucleus of an atom has very little energy and a lot of momentum (it changes the direction of the momentum of the electron, but it does not change the energy of the electron). It is far from being an actual (real) photon. So, the difference comes down to the fact that a short-lived real quantum is still *on average* on mass shell. And a disturbance may be far from it, even on average.

A real particle is a resonance. It has a sharper energy and hence a sharper frequency, since this is the natural frequency of the field (like the tuning fork). When it is short-lived it is always close (in time) to its creation and its annihilation. These are processes that change the frequency (“collapse” of the wave function) and are responsible for its frequency to be unclear. Otherwise, it lives for a long time and can be modelled as a single wave on the mass shell.

Fields can carry energy in disturbances and in resonances (quanta). A disturbance that is far off the mass shell decays away because it is non-resonant. A disturbance that is close to the mass shell decays away due to damping, not due to being off-resonance. A resonance on the mass shell has a tendency to stay. The consequence of this difference is that, statistically speaking, particles are more common than disturbances. Put differently, we measure particles, not disturbances. Even though the disturbances are an essential part of the interactions in the measurement process.

When we measure a particle, we may measure it at a time when it consists of many disturbances. However, these disturbances are linked. Summed up, they keep the characteristics of the particle. So, the interaction that supplied us with information about the particle does interact with the sum of disturbances. For example, when a particle is absorbed in the measurement process (e.g., a photon), all the linked disturbances are absorbed. They live on as new disturbances, becoming part of the particle that absorbed the photon.

Quanta and disturbances within a field are related to each other as they live in the same field. But there are many essential differences that govern their behaviour. The general popular literature leads us to believe that virtual particles are the same as real particles, with the difference being that virtual particles get pulled from the vacuum for a short time (based on the uncertainty principle). I hope that by now you have gained a much more subtle view on the matter and it is clear that virtual and real particles are very different beasts.

10.3.1 Summarizing

Table 10.1 gives us a summary of the differences between quanta (“real” particles) and disturbances (“virtual” particles).

So, we can conclude that an off-resonance disturbance in a field is the result of a phase shift caused by a source, with the consequence that the field becomes inhomogeneous. It is represented by the propagator, which describes such a disturbance as a bunch of waves that are off mass shell. These waves have a limited lifetime after which they are either recombined with the source or absorbed by another quantum.

A quantum on the other hand is a resonance in the field which can sustain itself for a long time and ideally is described by a single sine wave in a homogeneous field. It is represented by a wave equation that obeys the relativistic dispersion relation and is on mass shell.

When a quantum shifts phase, both the quantum wave and the gauge wave become a disturbance and are off mass shell. Summed up, they still represent the same energy, momentum, and mass as the original quantum and together they are on mass shell. When they recombine, they are the same quantum as before. When the gauge wave is absorbed by another particle, the original quantum becomes a sine wave again, but now with a different energy and momentum, and back on mass shell.

10.4 What Is an Electron Really?

When an electron produces phase shifts it disturbs the EM gauge field. It receives kick-backs from it, since this is the only way to make the EM field ring at an off-mass shell tone. As we have seen before, the electron spends its time as a combination of two disturbances: one in the electron field and one in the electromagnetic field (see Figs. 10.12).

Something similar goes for the photon itself (see fig. 10.13). It spends some time as a disturbance in the electromagnetic field and some time as a disturbance of the electron field. This disturbance has some regions with negative electric charge

Table 10.1 Summarizing the differences between real particles and virtual particles

Real particles	Virtual particles
Resonance	Disturbance
Tone of the tuning fork	Tuning fork forced to produce a different tone
No source necessary	Source (charged particle) disturbs the field to produce the “off tone”
Solution is a wave	Solution is a propagator
One wave	Wave packet
Long life/stable	Short life/quick to die
On mass shell	Off mass shell

Fig. 10.12 Feynman diagram of an electron producing a gauge photon and reabsorbing it again

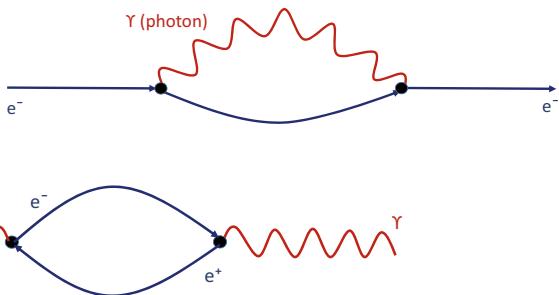


Fig. 10.13 Feynman diagram of a photon producing an electron–positron pair and recombining into a photon

and some with positive electric charge, but with total charge zero. The sum of the disturbances of the photon in the electron field always adds up to zero charge and zero mass. The photon can do this in all charged fields. For instance, it is sometimes a disturbance in the field of the muon (a heavy kind of electron belonging to a different particle family), and sometimes in the quark field.

A photon produces a particle–anti-particle pair in such a field. Here we use our previous insight that their net mass can be zero. Their waves are limited and they move in such a way that, together, they cause a net zero movement of the springs attached to the electron field. Or better, the net potential of the pair in the Higgs field remains zero. They also have a zero combined charge. Note that the electron–positron pair is in fact an electron that first moves forward in time, changes direction, and moves back in time again until it gets back to where it started. So, you can see this “pair of disturbances” as an electron loop!

All these effects can be properly calculated, but the calculation is so complicated that it would not be feasible to reason out all these effects conceptually. In order to get the right picture of these effects, they must be calculated. The calculation for the photon as being sometimes a disturbance in other (charged) fields gives exactly the outcome that the electron–positron pair has zero net mass and charge [Ref. 12].

When a photon can sometimes be a disturbance of the electron field, it basically polarizes the vacuum. The point is that when a photon is an electron–positron combination, there is somewhat more negative charge on one side of the photon path and somewhat more positive charge on the other side of the path. In general, this will change randomly so there will be no net charge effect and the photon will not disturb its own field. However, when there is another charge in the neighbourhood, say a negative charge, then the electron–positron combination will be biased in its direction. The positron will want to be closer to the negative charge (as it is positive itself), while the electron will want to be further away. We will use this process later on when we discuss vacuum polarization and a related infinity problem of the theory.

So now we can combine these disturbances into a picture of what an electron really is (see Fig. 10.14). The electron is made up of a whole bunch of disturbances.

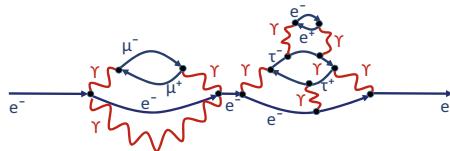


Fig. 10.14 An electron pictured as a bunch of disturbances in the photon field, which in turn has disturbances in other charged fields such as muon (μ) or tau (τ) fields

The funny part is that an electron sometimes gets to be a little bit quark + anti-quark, or a little bit muon + anti-muon.

We can go on with this process, in which, e.g., a tau disturbance can disturb the photon field, and that photon disturbance may in turn disturb the electron field or yet other fields and so on. You can imagine that each such branch into yet different disturbances spreads the energy over them. When the energy gets spread out, the energy for any such individual disturbance gets less. The more branched the disturbances get, the less energy will be available for each individual disturbance. Consequently, these disturbances become further and further off mass shell and less important in the calculation of their effects. In general, when calculations are made, effects are often only included a few branches deep.

The sum of all these effects at any one time is equal to the mass and charge of an electron (as seen from some distance). And so, an electron moves through the world. For many general purposes the electron can still be modelled as a single wave with one momentum, but in reality, it is a continuous melting pot of disturbances. This melting pot is sometimes called a virtual particle cloud. Only when a particle does not connect to other fields can it be a nice sine wave. Otherwise, it will disturb and get disturbed all the time.

The picture of elementary particles we create here may seem strange, and it becomes unclear how this mechanism of disturbances manages to keep track of the energy and momentum (not losing some on the way), but the calculations that are done this way agree with measurements to an incredible degree of accuracy.

It becomes especially interesting when the momentum of an electron is very high. When we have two such electrons bumping into each other, their combined energy may be sufficient to create, not a disturbance, but a real quantum. For instance, the two electrons may exchange a photon disturbance of very high energy, and this photon may then split into an electron–positron pair. Supposing that its energy is enough to produce not just the pair as a disturbance, but to produce two on-mass shell quanta, these quanta will be more stable than the disturbances and will fly off with a life of their own. So, two high energy electrons bumping into each other may lead to three electrons and a positron flying off. Hence, disturbances can become actual quanta when there is enough energy.

10.5 How Do Virtual Particles Create a Force?

We saw before that a force was created by charged particles on each other simply by bending the path of each other's waves. As explained, this is a consequence of an exchange of potential by the gauge wave that extends or shortens the wavelength (i.e., momentum).

In the present chapter we have discovered that such a gauge wave, transporting the potential difference, is really a virtual particle. In general, we have seen that a particle such as an electron creates many virtual photons and propagates through the world as an off-shell electron + a bunch of virtual photons and other disturbances. We can call this a cloud of disturbances. An interaction between two such clouds cannot be readily understood conceptually and really has to be calculated.

Nevertheless, we are not going to give up! Let's see how far we can get in improving our conceptual picture. For the discussion that follows, we limit the cloud of disturbances to an off-shell electron and a bunch of virtual photons, hence leaving the other disturbances out of consideration. This is good enough to get an idea of what is going on. In this picture, the total energy and momentum of the electron + cloud is that of the “classical electron”, so the cloud of virtual photons carries a part of that momentum and energy.

What would happen if such an electron were to meet another particle? That other particle also has a bunch of virtual photons around it. Their clouds would already start to mix at some distance. Since the virtual photons get created and reabsorbed all the time, they will inevitably start to absorb each other's virtual photons. From here on let's distinguish between electrons of equal charge and those of opposite charge (e.g., when an electron meets a positron).

10.5.1 Electrons of Equal Charge

Suppose the two electrons come flying towards each other from far apart. This means that they have opposite momentum, as do their virtual photon clouds since they carry part of that momentum (see Fig. 10.15). When electron 1 absorbs a virtual photon of electron 2, this means that it absorbs a little bit of opposite momentum. When they are still far apart this is not much, but it leads to a small decline in its momentum. The photons carry not only momentum, but also a spring strength (potential). For equal charges we have seen that potentials have to be added. This means that the spring strength that is carried by a virtual photon of electron 2 will add to the spring strength of electron 1 when it is absorbed by it. The other way around, things work in the same way: when electron 2 absorbs a virtual photon of electron 1, it will absorb some opposite momentum and additional spring strength. Hence, it will decrease its momentum and increase its spring strength. When the two particles get closer, they start to absorb more of each other's virtual photons. The result is that the effect gets

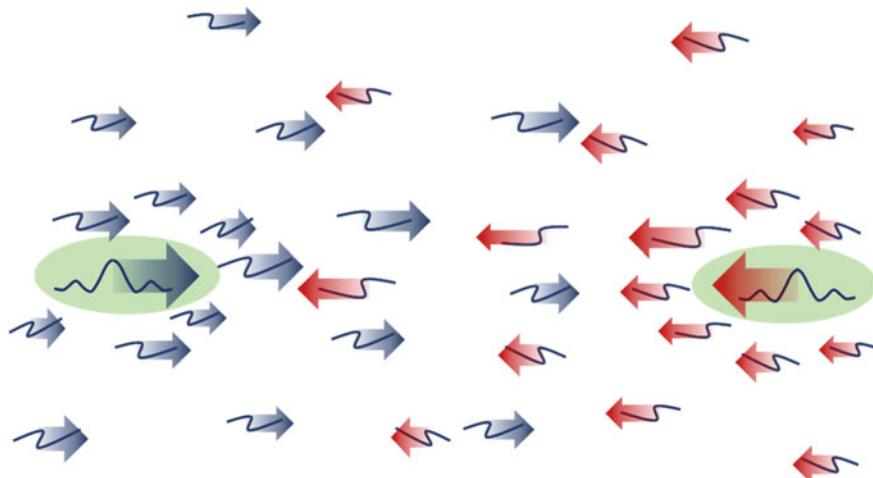


Fig. 10.15 Two clouds of virtual photons. Each electron shares the same colour with the virtual photons belonging to it. Both electrons have the same charge (blue waves), but their momentum is opposite (red and blue arrows). The various virtual photons are depicted by long and short arrows representing smaller and larger shares of the momentum of the cloud. Some arrows are slim and some are fat, which indicates the share of the cloud's spring strength that they carry. All this means that each virtual photon can carry a different share of the total energy and momentum. These differences will average out when many are absorbed by the other particle

stronger: the momentum of each electron will decrease further and its spring strength will increase.

Another way to view this is by thinking of the field strength of the electrons. Basically, it is the cloud of virtual photons that constitutes the field of an electron. When two electrons come closer, their spring strengths will increase. This higher spring strength is carried in part by their virtual photon clouds. So, the spring strength (i.e., field strength) available in these clouds is higher when the electrons are closer. This is logical! Classically, when you have two electrons close together, their charges add up and the electric field around them is twice as strong. This signals the increased spring strength carried by the virtual photons.

The cloud of virtual photons contains momentum and potential (spring strength). We can view this as a momentum density and potential density spread over space and propagating through it. As we have seen we may summarize the behaviour of this cloud as a single on-shell wave. However, when two clouds start to interfere, both waves change. When they change, they are no longer a perfect sine wave, which implies that they both go off-shell. The absorption of each other's virtual photons can be viewed as an exchange of virtual photons changing both waves. How should we connect these two views?

We can compare this situation to that of a gas in a bottle. There are trillions of individual gas molecules hammering the wall of the bottle. Added up, this leads to a pressure on the wall. The same goes for the temperature of the gas. When the

temperature rises, the molecules start hammering harder on the wall. Consequently, the pressure increases. So, the hammering of the molecules always adds up to the same simple relation between pressure and temperature. In a similar way we can view the cloud of disturbances, whose individual momenta and energies always add up to the same simple relation between momentum and energy (on shell). That relation can be modelled using a wave. This illustrates that the wave picture of an electron can be seen as the statistical sum of the cloud of disturbances that the electron really consists of.

When we allow two bottles of the same gas to mix together, but with each at a different temperature and pressure, they start to mix and the total temperature and pressure will quickly equalize until the difference is zero. Similarly, two electron clouds mix until the difference in momentum is zero. However, the comparison between a gas and an electron stops here: the mixing between the electron clouds can continue afterwards and the electrons gain momentum again, each in the opposite direction. This causes the path to bend in such a way that the electrons get closer together until the momentum is zero, but then fly off again. This effect can also be understood through the concept of “potential”. In the gas mixing process, everything gets equalized, but in the process that mixes electron clouds, the potential of two approaching equally charged electrons does not get equalized: it increases instead. Nevertheless, the relation between the total energy of the system, the potential energy, and the momentum (kinetic energy) is a simple relation that sums up all cloud behaviour. So, let’s try to understand the process of mixing of electron clouds using a wave picture.

The increase in spring strength (potential) can be understood using Figs. 7.3, 7.4, and 7.5. These show the wave of the spring strength, or the mass wave. When the spring strength increases, the potential narrows and the energy goes up. This means that the mass energy (or potential energy) goes up. Put differently, there is more energy in the springs of the wave. At the same time this means that the amplitude goes down a little (combine Figs. 7.3 and 7.5).

The total energy in the system of two electrons does not change, but the energy in the springs increases. We can use Fig. 9.15 to see that this increase in spring energy (or equivalently spring frequency) must lead to a decline in wave energy (frequency). The decline in wave frequency must lead to a decline in velocity and an equal increase in wavelength. We can summarize the whole process in a wave picture, as shown in Fig. 10.16. There we see that the exchange of virtual photons increases the spring strengths and wavelengths of both electrons. The closer the electrons get, the stronger the field and the more virtual photons get exchanged. From this we can conclude that the waves of both electrons are more extended when they get closer. Remembering that we have to consider all possible paths the electron waves can take, some waves will go closer to the other electron and some further away. The closer waves will be longer while those further away will be shorter. Using the path integral from Sect. 9.4 (“how does symmetry create a force”), we can understand how this leads to a repulsive force. However, we used a simplified picture to get to this result. In quantum field theory that is not exactly the way the *calculation* is done (see Sect. 10.6).

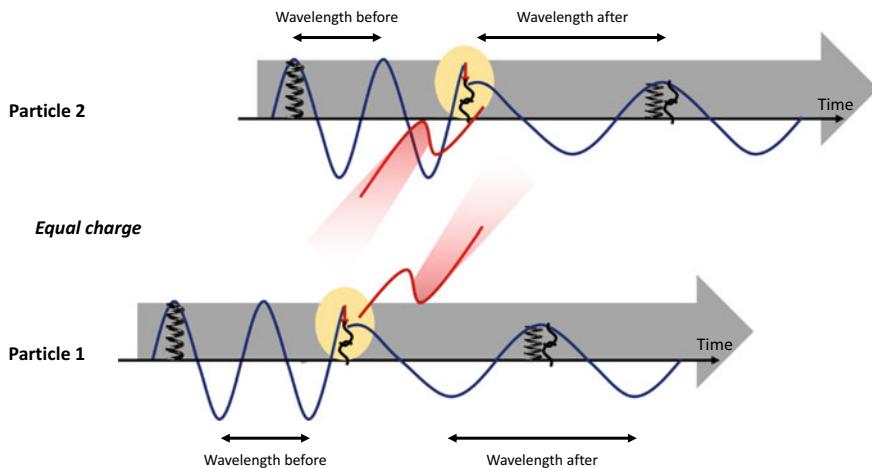


Fig. 10.16 Exchange of virtual photons in the wave picture. Equal charges increase each other's spring strength and each other's wavelength

10.5.2 An Electron and a Positron

What happens when the two particles have opposite charges? When the two particles fly in from far away, they will start to absorb each other's virtual photons (see Fig. 10.17). Suppose the electron absorbs a virtual photon from the positron. The spring strength of an opposite charge will have to be subtracted rather than added!

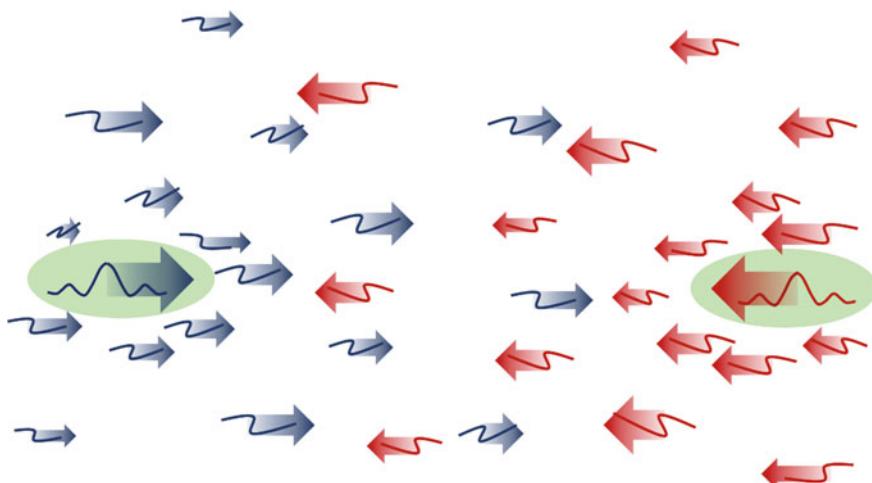


Fig. 10.17 The virtual clouds of an electron and positron. The two particles have opposite charges (red and blue waves) and they also have opposite momenta (red and blue arrows)

Earlier we compared the opposite charge with an anti-particle (see Fig. 10.7). This picture made it clear that the potential is absorbed in the opposite way and must be subtracted. That means that the spring strength of each particle will be reduced when they get closer and the field gets stronger. Consequently, the energy in the springs will also be reduced.

The total energy of the electron + positron system remains the same. So, we conclude that, when the spring energy is reduced, the momentum energy must increase. That means that the opposite momenta of the two particles does not cancel out but has to be added when they absorb each other's virtual photons, as if the momentum carried by a positron's virtual photon were absorbed in a reverse manner by the electron, just like the spring strength. Consequently, when the two particles close in, their momenta increase.

Again, we can view the two clouds from a wave perspective. Each particle plus its cloud is then modelled as a single on-shell wave before they interact. When the interaction commences, they have to go off-shell since the interaction starts to change the waves. This will last until the interaction is over. This time the decrease in spring strength translates into a wider potential, a lower spring energy (frequency), and a slightly higher amplitude (see Figs. 7.3, 7.4, and 7.5). However, *it is not the rest-mass spring that gets relaxed*. This process does not relate to Higgs. Suppose the particles are carried a great distance apart. In that case, the spring strength increases and the energy in the springs grows. So, when the particles are carried a great distance apart, we build up the potential energy. It is the spring strength of this potential energy that gets relaxed when the particles fly towards each other (see Fig. 10.18). Since the momentum has to go up, the wavelength gets smaller. So, when the particles get closer, the wavelength shortens. Looking again at the many different paths the waves could take, waves closer to the other particle will get shortened while waves further

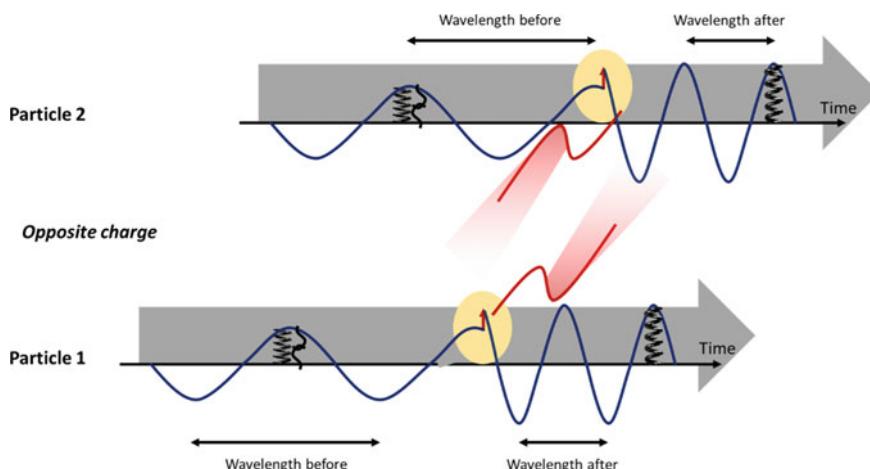


Fig. 10.18 Exchange of virtual photons in the wave picture. Opposite charges decrease each other's spring strength and each other's wavelength

away will get longer. Consequently, we can use the path integral again to understand how this leads to an attractive force.

We can also look at the field strength. When an electron and a positron get close together (not yet annihilating!), they neutralize each other's field. In the neighbourhood of the pair, the electric field quickly drops. This signifies the fact that there is not much spring strength left in the remaining virtual photons. It has all been converted to momentum.

10.5.3 Conclusion

We have seen how an electron with a cloud of virtual photons can be considered as a set of field densities in the photon field and the electron field. We discussed how virtual photons get absorbed and exchanged and we connected that view to field densities. We simplified the field densities into a single wave that can be seen as the sum of the cloud of disturbances. Using this wave picture, we could understand how the interaction shortens or extends the wavelength. From there on, the path integral does all the work to show us how the paths of interfering waves get bent, thus giving us the notion of a force.

The potential difference that is carried by the (cloud of) virtual photons is mathematically formalized into the so-called vector potential (\mathbf{A}). This potential describes both the electric field and the magnetic field. The combination of both into one potential expresses the fact that electric and magnetic fields are two sides of the same coin. Whether you experience an electric field or a magnetic field depends on your motion relative to the charged particles that source the field. Likewise, if two charged particles do not move with respect to each other, their mutual force will be purely electric, but if they do move, it will be a mixture of electric and magnetic. So, we understand that they really are the same phenomenon: the result of the electromagnetic interaction.

But you will argue that the magnetic interaction has a number of different characteristics compared to the electric interaction. This is true, but these differences are caused by the relative motion of the relevant charged particles. Take for instance the following situation: place a positron exactly in the middle between two electrons. It will experience no net force as both electrons will pull it with equal force in opposite directions. But if the positron is moving to the left, just at the moment it is in the middle between the electrons, it will feel a force! This is due to the fact that it experiences the virtual photon coming from the right as less energetic compared to the photon from the left. This is a consequence of the fact that the energy of the motion is added to that of the photon when it is coming head on (from the left) and it is subtracted when the photon coming from the right has to catch up with the moving positron. We can compare this to the Doppler effect. It reduces the frequency of the photon coming from the right, while increasing the frequency of the photon coming from the left. Consequently, the force experienced by two charges moving with respect to each other is different than the force experienced by two charges not moving with respect to each other.

10.6 Path Integral Revisited

The path integral we discussed before was the path integral as defined in quantum mechanics. It is related to the path of a particle. By now we know that we are dealing with a complicated interplay of waves and disturbances, rather than a “particle”. The path from A to B no longer involves a single wave. In quantum field theory, it involves all possible Feynman diagrams along the way, including a wide variety of interactions, self-interactions, and phase changes. And all possible variations have to be included. So how to formulate the path integral in this new landscape?

Classically, we use two types of energy in describing forces: (1) momentum or kinetic (T) and (2) potential (V). But as we have seen, V is a consequence of the virtual particles that swarm around a source (e.g., an electron). We have seen how T and V get exchanged. The behaviour of field quanta is determined by this exchange. In fact, T and V are highly connected: in a closed system (where the total energy does not change), T has to go up when V declines and vice versa. Consequently, we can define a single quantity that summarizes this exchange: the action. In field theory we have to consider both V and T to be energy *densities* of the field.

The exchange between V and T is determined by the phase shift. We saw that when a phase change occurs, a potential is carried away by the gauge wave. This potential being carried away signals the exchange between V and T : potential is added at the cost of velocity or velocity is added at the cost of potential. So, when we work with the action, instead of T and V , it is the change in the action that is determined by the phase change of the quanta. It summarizes the whole process of exchanging virtual bosons and exchanging V with T and vice versa. The change in the action is the quantity that says how fast V is exchanged with T . So, a large change in the action implies a fast exchange of potential energy V with kinetic energy T (or vice versa).

But wait a moment! How does all that relate to shortening the wavelength and hence bending the path, as in the path integral? When we use a single wavelength to model the (average) behaviour of a particle, the wavelength is related to the momentum ($P = 1/\lambda$). So, the kinetic energy T (which is determined by the momentum) is related to the wavelength. At the same time, the potential carried by the gauge wave is related to V . So basically, the picture of a changing wavelength agrees with the picture of exchanging potential energy (V) for kinetic energy (T). In fact, we can relate the phase of the wave directly to the action:

$$\text{Phase} = (\text{action}/\hbar) \times 2\pi$$

which makes Planck's constant \hbar a quantum of action. So, \hbar signifies an amount of action related to 360 degrees ($=2\pi$) phase. The first conclusion to draw from this formula is that a stationary action equals a stationary phase. So, when the action does not change, the phase does not change. This happens when the potential is constant and the kinetic energy is constant: there is no net exchange of potential for kinetic energy. Clearly, a particle in such a situation may undergo many phase shifts, but it will re-absorb them. A constant field also equals a constant potential. The particle

will absorb the virtual photons of that field, but everywhere in the same amount. Hence, the action is still 0. It is only when the particle absorbs different numbers of virtual photons in different areas of space (hence, an inhomogeneous field) that its wave will start to change, and consequently, its phase will also start to change and we will have action!

In an inhomogeneous field, the phase will change a little with every small step along the path. Hence, the action will change with every step. We can formulate this roughly as follows:

$$\text{Change in phase} = 2\pi \times (1/h) \times \text{change in action}$$

Change in action = the exchange of V with T in a
step along the path in space–time

So, the *total* change in action along a path is the *sum* (or rather the *integral*) over all steps in space–time of the exchange of V with T x step. And the total phase change equals the total change in action $\times 2\pi \times (1/h)$.

Summarizing, to calculate the path integral we need to adopt the following process:

1. Identify all possible space–time paths from A to B.
2. Calculate V in the field exchanged with T along all paths, using all possible interactions.
3. Find the total phase for each path by summing up the phase change in each step along the path.
4. Add the phases at the end of each path (superposition).
5. The end result is the probability of going from A to B.

We find that the highest contribution to go from A to B comes from the path with the least phase change (least action). The path with the least phase change is the path that interferes most positively with nearby paths. It is the shortest path. So far, we are doing this much as we would in quantum mechanics and we still have the same result as before: paths that bend.

The difference comes when the exchange of V with T is determined by the superposition of field quanta instead of a continuous field. The quantization of the field (second quantization) is specific to quantum field theory. This means that we have to add up all field configurations, including, e.g., photons splitting into electrons, recombining, and creating leptons of a different flavour. Along each path we find a large number of such vertices of bosons connecting to the particle. Therefore, we can calculate the exchange of V with T by finding all possible particles, interactions, and vertices, i.e., by finding all relevant Feynman diagrams along the path.

In conclusion, we can say that the path integral in quantum field theory is calculated using all possible Feynman diagrams along a path rather than a wave along a path. This is very abstract, but the result is extremely accurate. The action of all the virtual particles (gauge waves carrying potential away) still bends the path of the particle wave. Even though “the wave” is actually some combination of field densities of numerous disturbances.

10.7 Fluctuating Fields

Quantum fields are constantly fluctuating. You can compare this to the sea. Even when there are no clear waves present, the surface of the sea is always moving in a random manner. It is fluctuating. This is something fields do as well. There are always disturbances. They die out so they do not carry energy from A to B. You can view this just as well using the uncertainty principle. During a time ΔT , an energy ΔE can exist, as long as it disappears after that time. This means that a packet of waves (a disturbance) can exist that dies out. This disturbance has to end up with the same energy and momentum it started with: 0. So it can exist, but leaves no trace.

The disturbances that can spontaneously exist for a short time in the vacuum can be shown in a typical type of Feynman diagram, called a vacuum diagram (see Fig. 10.19). Such a diagram is characterized by the fact that it does not show external lines!

You might wonder whether such fluctuations really exist? If they do not leave any trace, how can we know? Well, there is an interesting experiment that indicates their existence. This experiment is based on the Casimir effect.

10.7.1 Casimir effect

The Casimir effect can be understood to give proof of the existence of virtual particles in fluctuating fields. The idea is to take three uncharged metal plates and put them close together in parallel. Classically, these plates should feel no force, since there is no charge and therefore no electromagnetic field present.

However, when we assume the existence of disturbances in the field, the following can happen. The plates repel electromagnetic waves above a certain wavelength. When the wavelength gets too short, it can pass through the plate and leak away. But longer wavelengths can only exist between the plates when they fit between the plates. Hence, the scale of wavelengths that can exert a pressure on the plates is limited to the interval between the shortest wave to be repelled and the longest wave to fit between the plates.

So, we have three such plates (see Fig. 10.20). Now suppose we put plate 2 very close to plate 1. Then the space between plates 1 and 2 can contain a smaller interval of wavelengths compared to the space between plates 2 and 3. This means that when virtual particles do exist, there would be fewer of them available on the left-hand side of plate 2 at any given time than on the right-hand side of plate 2. Effectively,

Fig. 10.19 Feynman vacuum diagram



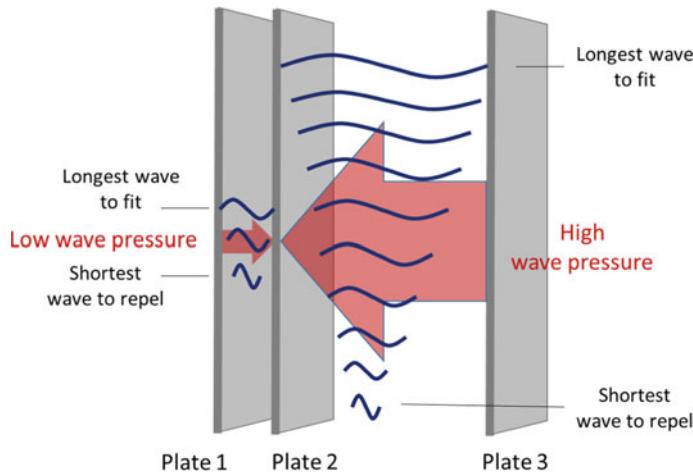


Fig. 10.20 Casimir pressure on a metal plate

the total pressure of the virtual particles on plate 2 will be bigger from the right than from the left. The net pressure pushes plate 2 towards plate 1.

This effect is extremely small, but could be measured. It takes a separation between plates 1 and 2 of about 10 nm in order to get a Casimir pressure of about 1 atmosphere on plate 2. This is a distance approximately 100 times the size of an atom. The number of virtual particles gets significantly lower when the plates are very close together.

A surprising comparison is that we know this effect from ships in the water as well! When two ships sit next to each other in the water, but with some space between them, they will start to move towards each other. The reason is similar: there can be fewer waves between the ships than around them. The reason for that is that waves longer than the distance between the ships will not fit between them. So, the pressure of the waves outside the two ships is greater than the pressure of the waves between the ships and the waves will start to push the ships towards each other.

This effect has been taken as proof of the existence of fluctuations in a field if we accept that only virtual particles could be responsible for the pressure experienced. Hendrik Casimir, a twentieth century Dutch physicist, originally developed the theory to describe the force. There is also an alternative explanation based on the van der Waals force between the plates [ref. 67]. This explanation does not rely on vacuum fluctuations. Hence, the question remains open!

10.8 The Arrow of Time

Let's take a deeper look at the idea that anti-particles are particles that travel back in time. Does that create problems for causality? Causality is sensitive to the direction of

time. After all, cause precedes effect in order to have causality. But what is causality really?

Suppose I have a single particle A coming from direction 1, hitting another particle B, deflecting and moving off in direction 2. For this process we could easily turn around the arrow of time. In that case we see particle A moving in from direction 2, hitting particle B, and moving off in direction 1. So, this process can be turned around without any problem. Causality still exists, but now in the other time direction.

Now let's up the stakes a bit and look at a snooker-like process where a ball hits a bunch of balls on the table. In that case, when we turn back time, we would see a bunch of balls coming together exactly at the same time and hitting one ball that flies off while the other balls come to a complete standstill. Although in principle this would be possible, it is extremely unlikely. So now we see a preference for an arrow of time. Causality seems to work one way but not back. Why is that?

In the first example, the probability for the process to happen one way is exactly the same as the probability for it to happen in the other direction of time. The reason for that is that the number of ways this process can go in one direction of time is the same as the number of ways this process can go in the other direction of time.

But this is not so for the balls in the second example. One ball hitting a bunch of balls leaves open many possibilities since we are allowing the balls that move off to go with any speed in any direction. However, the other way around has far fewer possibilities. It takes all the balls to be moving in exactly one particular way if they are to hit the other ball and come to a standstill, and this is just one of many ways this process could otherwise go. Therefore, it becomes highly unlikely that exactly this one process will happen. However, note that the following is also very unlikely to happen: we hit the balls with one ball and each of the balls goes exactly in the direction of the holes at the sides and in the corners of the table and they each come to a standstill just before each hole (very frustrating). This won't happen either because it is so unlikely! So, the perception of an arrow of time is related to the probability of a process happening. When we impose many restrictions, the probability for a process to happen is very small. If we allow all variations of a process to happen, one of them is likely to happen. It is the same as buying a million lottery tickets instead of just buying one. The chances of winning something are much greater with a million tickets. If we hit the balls and we are content with any direction the balls can move off in, we will win every time. If we insist on just one specific way for these balls to move, we may have to try a million times before we actually see it happen.

Any system is likely to move from a situation that has a low probability of happening towards one that has a high probability of happening. This is the increase in entropy described in thermodynamics and it determines our perception of the arrow of time.

However, we may still ask why time goes one way instead of the other? Why this preference? If we say that matter moves in one direction in time and anti-matter is in fact matter that moves in the other direction, we have a simple answer to that. In the universe there appears to be more matter than anti-matter. Consequently, we are only meeting matter. Anti-matter quickly gets annihilated. And the arrow of time agrees with the direction matter goes in. Why is there more matter than anti-matter?

At this point, science does not have an answer. But we can always fantasize a little about what we do not know. One thing we do not know is how the big bang got started. Time and space are said to have been created at the big bang. Maybe we can assume a further symmetry, namely that the universe starting in one direction may have its counterpart in the other direction, just as throwing a stone in the water gives a wave in all directions, not just in one direction. In this fantasy there could exist an anti-matter universe going in the opposite direction of time. Who knows?

Suppose the universe consisted primarily of anti-matter. Then we would experience processes as developing in the other direction of time and we would see entropy amongst anti-matter develop in that direction. Now let's assume you are a spectator from our world entering (with some kind of protection!) the world of anti-matter. What would you see? You would experience eggs coming together, shattered glass turning into a wineglass, and so on. It would truly be a spectacle to witness that! However, for the anti-particles that this world is made off, the egg just falls to pieces and the glass shatters when it hits the ground.

So, the direction of time is simply that of the prevailing type of matter. Causality exists either way in time. It is just the large quantities of particles that make the probability that certain categories of processes will happen more likely than other categories. Take the category of all processes by which an egg could fall and break into pieces. This category is huge: there are so many ways for an egg to break. But the category of processes for pieces of an egg to come together and become a whole egg is very small: there is probably only one way of doing this. The probability of this happening is not 0, but it might as well be when we compare it with the unimaginably many other processes that may happen with the many pieces of eggshell and the fluids within.

But hold-on! This can't be it! Can't I use anti-matter to send a message to the past and break causality that way? Well, not really. The problem is that the rest of the world is still moving forward in time. Let's try it out!

Suppose I would like to know whether my favourite horse will win later today. I happen to have a lab that is able to produce positrons. So, what I do is I produce a large number of high energy photons and wait for positrons to come from the future so that they can be turned into electrons by interacting with photons (see Fig. 10.21). In the future, I will create the positrons if my horse wins. Otherwise, I will not. So, when I find a bunch of positrons meeting up with my photons, I know the horse will win.

Suppose I do get a bunch of positrons interacting with my photons. What really happens is that a substantial number of photons will split into electrons and positrons (as seen relative to my arrow of time). I carefully catch the positrons in a vacuum jar so that they can survive until later today (how's that for a message in a bottle!). If my horse wins, I will combine them with electrons to annihilate them into photons. If my horse does not win, I will not combine them with electrons. Huh? But when I have the positrons, I know the horse will win, so I must combine them with the electrons. However, I can still make a choice to combine or not to combine them!!

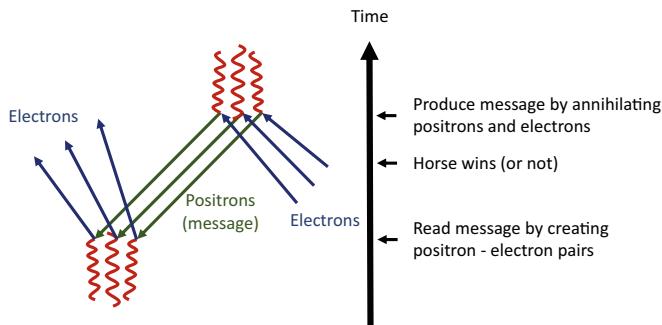


Fig. 10.21 How to receive a message from the future (or not?)

So, what happens when I do *not* combine them with electrons? They still exist, and they will annihilate at some later time. So, the fact that I received the positrons in the past does not say anything about how they were created in the future! And now we see that what matters is still our own arrow of time. I won't be able to send a message back in time, since the future still has to happen and it remains unclear how (and why!) the message will be created in the future.

But you may argue, what happens when you do not receive positrons at the beginning? Doesn't that say that the horse will certainly not win? Not really. For instance, I might find out that the horse wins and I would like to create the positrons, but I do not have positrons in the jar (in my arrow of time), so I cannot combine them with electrons. Here you start to see the problem. The only way to make electrons turn into positrons is by eliminating them. Consequently, *in our arrow of time*, we need to have those positrons ready. We need to nurture those positrons until (i.e., "until" in our direction of time) we are ready to annihilate them. It would be different if we could make the electrons turn into positrons by some other interactions or means, but we cannot.

So, it comes down to the fact that we cannot actively make matter turn into anti-matter! Consequently, we are not able to create a message to the past. This inability is not a lack of technology, it is impossible on principle. Hence, causality is preserved.

And so, we may conclude that causality exists in both directions of the arrow of time. The way processes will develop depends on the chances of the process happening, not so much on the arrow of time. Our perception of time is based on entropy growing in general. That is (by far) the most likely way for processes involving large numbers of particles to develop. Hence, large numbers of particles will make us experience time as going in one way, while a universe made of anti-matter would experience exactly the same but in the other direction of time. Only for simple processes involving one or two particles is the arrow of time perfectly reversible. The probability of such a process happening in one direction of time is

equal to the probability of it happening in the other direction. The only exception are weak interactions: they are not symmetric in time. We will get back to this later when we discuss so-called “CP violation” by the weak force.

Chapter 11

Renormalisation of Fearful Infinities

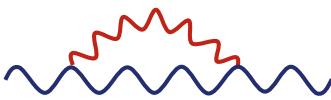


When we consider the landmark equation of QED, it contains the waves of charged particles and photons, it describes the springs that give mass to the particles, and it includes the interaction between particles and photons. In the previous section we saw, e.g., that an electron splits off some of its energy to create a virtual photon. How is that process described in the QED equation? It would require the mass springs to get less strong, for instance. But the equation does not describe any such thing. It only describes the perfect waves, the springs (that are not altered), and the interaction.

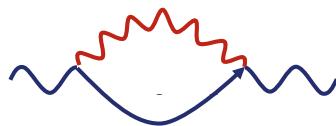
So, it seems we are missing something. But what is the problem with that? In order to calculate the probability for an electron to move through a field with which it interacts (such as the electromagnetic field), we need to add up all Feynman diagrams. Basically, the problem is that the number of virtual photons that is produced in the way described by the QED equation could be infinite, and there is nothing we can do about that. The electron just keeps producing gauge waves and these all have some energy, but we are not subtracting that energy from the electron in the QED equation. The result is that the electron plus its photon cloud would have a much higher mass than we observe. Under certain conditions, the electron mass could go to infinity! So, we must find a way to reduce the mass of the electron by the same amount as the energy that is taken away by a virtual photon.

The infinity problem is worsened the closer we look. Closer means with a smaller length interval and a shorter time interval. When we limit the interval, we get a higher uncertainty in momentum and energy. So, the closer we look, the more energy is involved in the process of producing virtual photons and the more the vacuum fluctuates around the electron. At the same time, the shorter the interval, the lower the probability for this to happen. We could choose to add the processes of many short periods together, but that would only add up to a mean value that is comparable with a larger scale measurement.

How to resolve this? One way to go is to limit the length and time scales. So, we would not look more closely than a certain scale. Basically, we cut off the processes that happen within a certain “smallest” scale. But what scale to choose? We could use



Original QED equation: The electron wave is a pure wave and the photon wave is a pure wave. Hence, the creation of a virtual photon involves adding its energy. But where does that energy come from?!



Corrected QED equation: The electron wave splits into a photon disturbance and an electron disturbance. Hence, a virtual photon is created at the cost of the energy of the original electron.

Fig. 11.1 What wave does QED describe?

the Planck scale, but there is no formal indication that we can put a limit anywhere. We will get back to this when we look at the renormalisation group.

Another solution involves adding extra terms to the QED equation. These terms are formally called “counter-terms”. The counter-terms take care of reducing the mass of the electron by the same amount as the energy put in the virtual photon. This seems a bit artificial (and in fact it is), but let’s see why this has a certain logic to it. The counter terms basically acknowledge and reflect the fact that the electron is not a perfect wave. So, the wave gets a counter-term that turns the perfect wave into a disturbance. What we are saying then is that, instead of taking the perfect electron wave as a basis and adding virtual photons, we now take the electron + photons as a basis and describe them as disturbances that together add up to an electron wave. Actually, just as we did when we were explaining what an electron really is (see Fig. 11.1).

The process of adding counter-terms to the QED equation is called “renormalization”. In this process the field wave gets rescaled and some of it becomes a disturbance in the photon field. When there are no interactions, this term is 0 and the wave is a simple wave. The interactions could also be so strong that it becomes impossible to create an electron, since it immediately interacts so strongly with the photon field that any attempt to create an electron immediately leads to numerous disturbances in the photon field. If the interaction is this strong, the counter-term cancels the wave. In general, this term spreads the disturbance over both fields, depending on the interaction strength.

The consequence is that the new QED equation describes a wave packet of disturbances in both fields. The more waves in the wave packet, the faster it dies away, giving birth to the next set of disturbances or, maybe for a short time, a single electron wave.

The counter-terms have consequences for the mass and interaction potentials. If there were no interaction, there would be no virtual particles around the electron and no self-energy impacting the mass. So, let’s take a look at the consequences for the mass and interaction potentials.

11.1 Renormalizing Mass

In the original QED equation, the mass was the mass that goes with the perfect wave and it was described by the mass potential (“mass spring”). When we add a cloud of virtual photons to this, the energy of the whole goes up and can even go to infinity. The energy in these self-energy processes equals mass, so the mass can also go to infinity.

In the corrected QED equation, the counter-term reduces the mass (see Fig. 11.2). It shifts the mass to the correct electron mass (as we measure it from a distance). However, there is nothing in nature that tells us how to arrive at the right correction term! Here is the weak spot of quantum field theory. Basically, we choose the correction term so that we arrive at the measured electron mass. So instead of taking the electron mass and adding virtual photons, we take the mass of the electron + photon cloud to be the electron mass (see Fig. 11.2).

An intriguing question is this: what is the remaining mass of the electron that is hidden in the photon cloud? Say we subtract that photon cloud and measure the mass of the remaining electron. How big is that? This is unclear since it depends on how closely we look! The closer we look, the more we penetrate the photon cloud and the more we see the hidden electron, but we will never get close enough to see the pure, naked electron. The mass of the naked electron is called its “bare mass”. The mass of the electron including its photon cloud is called the “dressed mass” of the electron, and such an electron is called a “dressed electron”. The electron is pictured as being dressed by its photons.

The mass shift in a field does not only happen in the vacuum. There are also examples of particles in a medium. For instance, an electron in a crystal interacts with the crystal ions and behaves as if it has an effective mass M^* that is higher compared to the mass of an electron in vacuum. We can study and confirm this difference in mass by taking the electron out of the crystal and putting it in a vacuum. The interactions in the crystal act like a potential that changes the electron wave as a whole and makes it more massive. So, the electron acts with greater inertia than the “bare” electron outside the medium.

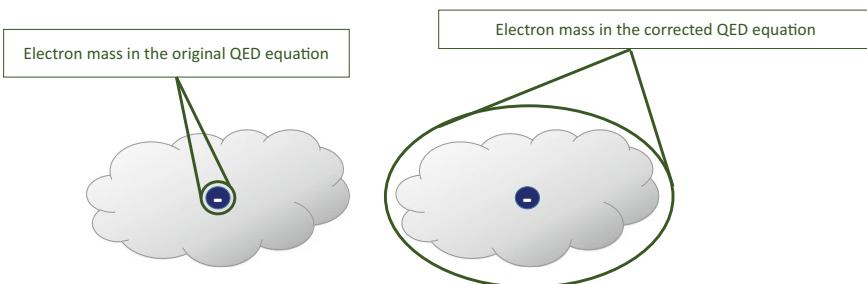


Fig. 11.2 Shifting the mass

The same happens to the electron in the interacting electromagnetic field. The interaction increases the mass of the bare electron to that of the electron we measure. The mass of the *bare* electron is caused by the interaction of the electron with the *Higgs field*. Both interactions have the same consequence: both fields act like springs attached to the electron wave and the potential they create changes the way the basic electron wave behaves. It changes the mass and the dispersion relation of the electron.

The only problem is that we cannot measure this effect as we did with the crystal, simply because we are unable to take the electron out of the vacuum, out of the electromagnetic field, or out of the Higgs field. All these fields are everywhere in space. So, there is no way to take the electron out of their influence and compare with the characteristics of the electron as they would have been outside these fields. Hence, we can only measure the fully dressed electron.

In conclusion, we can view the dressed mass of the electron as enhanced by the springs of the interaction with the electromagnetic field. Equivalently, we can view the dressed mass of the electron as its bare mass + the mass that is stored in the virtual photon cloud around the electron. If we consider the virtual photons to be like springs attached to the electron, then the energy of a virtual photon is just the energy stored in its spring connected to the electron. And so, we can see the similarity between the two views.

11.2 Renormalizing Charge

What happens to the interaction potential? For this we first need to see what effect the photon cloud has on the interaction potential. We saw before that a virtual photon can produce a virtual electron–positron pair which annihilates back into the virtual photon. We can view this as the virtual photon disturbance kicking back on the electron field and thereby disturbing it back. The photon can also disturb other charged fields. All this happens in the neighbourhood of the original electron. Since that electron is charged, we get a strange process called vacuum polarisation (see Fig. 11.3).

In this process the virtual electron–positron pair produced by the photon feels the charge of the original electron (feeling means interacting, so a virtual photon gets exchanged between them). The consequence is that the pair experiences a “force”: their paths get bent in the electromagnetic field of the original electron. Therefore, the positron gets bent towards the original electron and the electron gets bent away from it. This leads to a gathering of positive charge around the original electron and a ring of negative charge a little further away. You might say that the cloud of virtual positrons and electrons around the original electron screens off its charge but also produce new charge towards the outside world. The result is that the vacuum around the electron is not homogeneously neutral. There is more positive charge towards the centre and more negative charge towards the outskirts. We say that the vacuum gets polarized. This process is called “screening” of the charge in the centre.

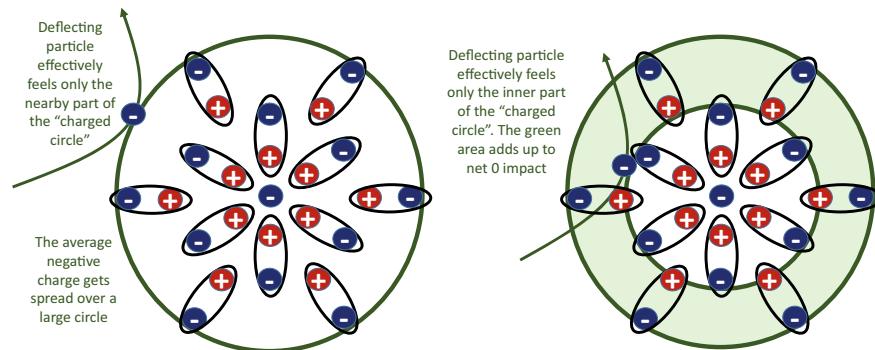


Fig. 11.3 Vacuum polarisation

Screening affects the charge distribution around the electron. It is no longer a point charge, but a “spread out” charge. From a certain distance it may still look like a point charge, but at close range, the effects of the charge distribution have an impact on the interaction with other particles. This effect can be measured as it results in shifts in the energy levels of the electron in, e.g., the hydrogen atom. These shifts are called Lamb shifts. The theoretical predictions of this process explain the measured energy shifts extremely well [Ref. 60, p. 196].

Since the charge gets “spread” over a whole region, the field looks weaker than it would if the vacuum did not create this effect. So, if the vacuum did not produce virtual electron–positron pairs, the spreading effect would not be there. In that case, we would see the bare electron charge. If you look at Fig. 11.3, you see that a deflecting particle experiences a charge, but the average charge is spread over the entire circle, so outside the circle, the deflected particle experiences a *lower* charge. Since in reality the “photon and pair cloud” has no clear boundary (there is no circle), this effect extends outwards.

On the other hand, if you could enter the circle and get closer to the original electron, the screening effect would be reduced and the charge actually experienced would go up exponentially. But you may argue, the charges further away will still affect you. So why does that not have an averaging effect? For such a charge distribution, you can calculate that roughly speaking the effects of the charges outside the inner circle (i.e., the green area) will cancel each other out. So, the net charge experienced is the total charge added inside the inner circle. The smaller the inner circle gets, the less screening and the more you will see the bare charge of the electron. However, when you reduce the radius of the inner circle to 0, the charge will go to infinity.

Consequently, we need to take this process into account. The correction of the QED equation aims to get the actual disturbances, the actual mass, and hence the actual charge into the formula. So instead of a single wave with charge and mass that can all go to infinity when the virtual particle cloud is considered, we now take the disturbed wave, shifted mass, and shifted charge into account. The result is that the

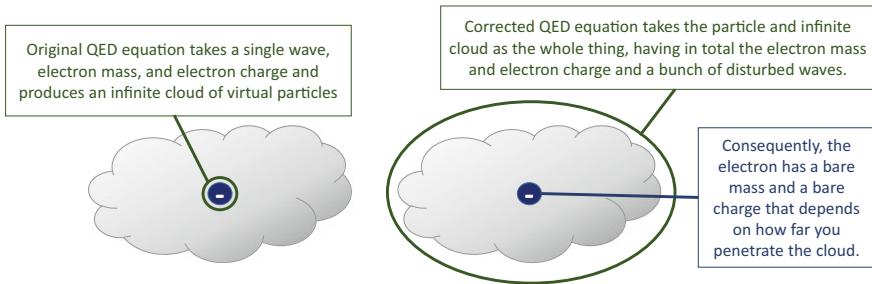


Fig. 11.4 Renormalisation: the corrected QED equation equates the bare electron + virtual cloud to the electron mass and charge

virtual particle cloud is included in the corrected QED equation to make the whole thing equal to the mass and charge we measure (see Fig. 11.4).

The charge of the electron plus its cloud is called the “dressed charge” of the electron and the charge of the electron without the cloud is called its “bare charge”.

Something interesting happens when we apply a strong electromagnetic field. Suppose we have a virtual photon producing an electron—positron pair. In a strong field these two can be separated from each other. But when they are just disturbances, what will happen to them? When the field is weak, the only possibility is that they eventually recombine into a virtual photon. But when the field is strong, they can absorb virtual photons from the field and gain enough energy to become a real electron and a real positron. This is pair production in a strong electromagnetic field.

A vacuum fluctuation can also be the source. In that case, the virtual pair is assumed to exist for a very short while (within the uncertainty limit). But when the particles absorb energy from the virtual photons in the field, they can become real. This requires the ability to gain an energy of two electron masses (electron + positron!) out of the field. So, the field must be very strong to do such a thing.

11.3 Renormalisation Group

The renormalisation group is not a group like those we discussed in the section about symmetry groups. The goal of renormalisation group analysis is to discover how the coupling constants change with the scale of interest. So, for instance, when we close in on the electron, we find less screening between us and the charge and the charge goes up. Hence, the coupling “constant” goes up.

How do we get closer to the electron? Basically, we are probing a shorter length scale. To do that we need a shorter wavelength. The wavelength must be of the order of the length scale to be able to “see” at that scale. However, a shorter wavelength implies a higher momentum.

Another way to see that is when we make a Fourier analysis of a signal. This means that we describe the signal using a set of waves of all frequencies. When we

cut off the higher frequencies, we lose detail in the signal. For instance, if we make a Fourier analysis of a picture, cutting off the high frequencies and translating the waves back into a picture, that picture will be blurry: the finer details will have gone!

So, the higher the momentum and frequency we probe the electron with, the greater the level of detail and the shorter the length scales investigated. This way we close in on the electron, and the result is that we measure a higher charge. The strength of the electromagnetic coupling is described by α , the fine structure constant. At normal energies it has a (dimensionless) value of about 1/137. At energies close to 81 GeV (hence closing in on the electron), it has a value of about 1/128, so slightly higher [Ref. 68]. How α evolves at still higher energies is not known, but it is assumed to go to infinity. At extremely high energies, quantum field theory no longer applies, due to the structure of the vacuum (we will get back to this in the last chapter). So, we are talking about very high energies, that are not usually available. However, there was a time when this was normal: the very early stages of the universe. In those days, the electromagnetic coupling strength may very well have been higher.

The weak force and the strong force also change in strength when we look at significantly higher energies. The coupling constant of the strong force actually *decreases* at very high energies! We will see later why that is. It does mean, though, that the strong force does not go to infinity at very close range! So, the accompanying theory, QCD, does not suffer from that infinity problem.

So, what happens when the length scale gets really short? Let's first compare with the field theory of a crystal. In a crystal there are no problems with infinities! The reason is that there is a shortest length scale: the scale of the crystal lattice. When we get to length scales on the order of the distance between two atoms in the lattice, the field approach starts to fail. This is logical, since it is impossible to define phonons at a scale between two atoms: there is nothing there that could carry the phonons. So, the field ceases to be homogeneous at atomic scales.

What does that tell us about the vacuum? If the vacuum were homogeneous even at scales below the Planck length, we would have a serious problem with infinities. However, it is not far-fetched to assume that the vacuum itself may have a structure at some scale. In that case, the field theory will not apply there and it will have to be replaced by some other theory at that scale. Field theory would still be applicable, but only at higher length scales, just as happens with the crystal.

Even though this is all speculative, the conclusion we can draw is that we simply do not know what happens at such a length scale, and until we do, we cannot say that field theory is wrong because of these infinities. We can only say that field theory is OK at length scales we do know and that we cannot use it at energy and momentum scales that correspond to shorter length scales.

So, what theories could there be for shorter length scales? Actually, there are no theories for that. There are a number of hypotheses, but none of these could be either verified or falsified by experiment so far. Hence, they will remain hypotheses until we do so. Some of them have been worked out mathematically in great detail, but they all predict things we cannot measure. Examples are string theory, quantum loop gravity, and causal dynamical triangulations. These hypotheses all include gravity, which we have not yet been able to combine with field theory.

Chapter 12

Is the Cat Dead or Alive? How Quantum Decoherence ‘Digitized’ the Universe



So far, we have explored a world of waves and virtual particles. But that is not what we see around us. What we experience when we observe the world are concrete things which are in a definite position or a definite state. In Chaps. 3 and 4 we discussed why we do not see the wave behaviour around us and how interactions create information about the position of a wave. Nevertheless, would it not be possible to bring the quantum world of interference between waves to an observable level? And how could we understand better why we see wave—particle duality in the double slit experiment? For example, how can we understand better what happens when the wave hits the detector? The wave may be extended across space but we nevertheless measure the particle at one position. In Sect. 4.3 we discussed the fact that interactions change the momentum of the wave, leading to a wave packet that has a clearer position. Now it is time to deepen our understanding of this concept.

Schrödinger was a quantum physicist who tried to make the problem of wave—particle duality explicit by means of a thought experiment, called Schrödinger’s cat experiment. He proposed (but only in thought!) to put a cat in a box with a device that measures a radioactive atom. The atom is in a quantum superposition of two possible states: decayed or not decayed. If the atom decayed, the measurement device would release a deadly gas that would kill the cat. When the box was closed, the cat was still alive, but as of that moment we don’t know what state the cat is in until we open the box and take a look. So as long as we do not look, is the cat alive? Is it dead? Or is it in a superposition of alive and dead? After all, the atom determines the state of the cat and the atom is in a superposition of the two states. We might try to imagine how this system could become a big pool of waves that inhabits all possible combinations of alive and dead. But somehow it is clear that this will not really be the case. Schrödinger wanted to tickle our minds by making us feel this explicit paradox. The problem is that this story has no satisfying end because there is no explanation of why we experience a concrete world where things are clearly alive or clearly dead and not in some observable superposition of the two.

Our experience that things are in a definite state can be seen as a classical “digitized” world. In that world things are either 0 or 1, heads or tails, dead or alive. Of

course, the world is not black and white, but full of grey scales. But even if you have fifty states of grey, they are not in superposition. We generally observe things in one of those states. Let's see if we can understand why that is. And how the world of waves and virtual particles transitions to the classical world.

12.1 Quantum De-Coherence and “Collapse of the Wave Function”

We have built up a picture of a particle as a cloud of virtual particles or rather disturbances. That cloud basically adds up to the total field quantum. Consequently, it is this cloud that (summed up) represents the wave function of the particle. The wave sometimes behaves as a particle. How can we see that in the picture of a cloud of disturbances rather than a simple wave?

Let's first look again at the absorption of a field quantum. We discussed this before in Sect. 8.3. Back then we said that the excitation of the field (a quantum) cannot exist when part of it gets absorbed, and consequently the whole thing must be absorbed. This is still right, but we can now make this a bit more precise.

We saw what happened in the exchange of virtual photons: they are produced by a quantum such as an electron that splits up into a virtual photon and a virtual electron. The virtual photon can transfer energy and momentum to another quantum without the electron ceasing to exist. To do that, the second quantum needs to absorb the virtual photon. A real photon can also get absorbed by a charged quantum. How can the photon actually be absorbed in one distinguishable position? How should we view this absorption process?

First of all, we need to realise there is a difference between the mass energy and the kinetic energy (see Sect. 6.2). The mass energy (at rest) cannot be shared without destroying the field quantum. As we saw before (see Sect. 8.3), for fermions there is no known process that would do this except if a fermion meets its anti-particle and the quanta cease to exist (while producing high energy photons). However, the kinetic energy can be shared, leading to changes in momentum. Massless particles (such as photons) can be absorbed, but again, the quantum has to get absorbed as a whole. For a photon to only exchange momentum primarily means that it gets deflected.

Secondly, since we are not talking about a single wave or a single quantum, but about a cloud of disturbances, we have the following situation: take a photon that gets absorbed in a detector. The photon has a cloud of virtual electron—positron (and other!) pairs around it. When it gets into the neighbourhood of an atom of the detector, the cloud will start to interfere with the cloud of virtual photons around an electron in the atom. Some of the virtual pairs may recombine into a virtual photon that gets absorbed by the electron, and the electron will thus gain energy. This process is highly statistical and soon becomes irreversible.

To see this, compare this process to the mixing of gases in thermodynamics (see Fig. 12.1). Take two gases, one blue and one red, and put them next to each other.

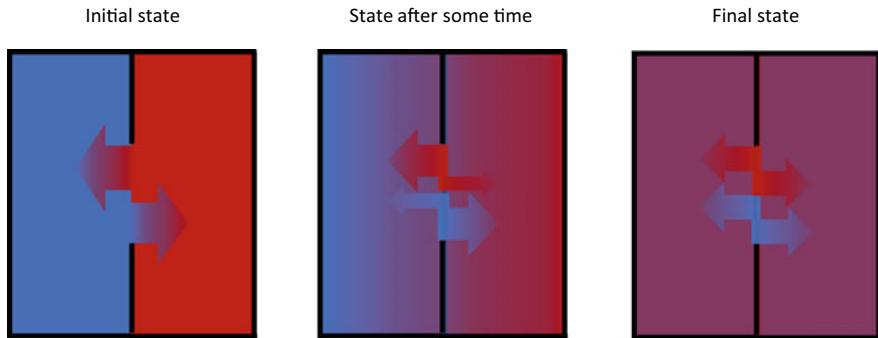


Fig. 12.1 Mixing a red and a blue gas. On the left the dividing wall has just been opened. There are no red particles on the left, so the arrow shows that they can only move to the left and vice versa for blue. In the middle, there are enough red particles on the left for some to move back to the right. But there is no equilibrium: the arrow to the right is still smaller than the arrow to the left. In the right-hand picture, there is equilibrium and the two gases have mixed completely. There is no practical probability of returning to the initial state on the left

Each gas represents a virtual particle cloud of one particle. So, we have a situation similar to two particles with their clouds sitting next to each other. Now we open the separation between the two gases. What happens? There is always some movement in the gases so they will start to mix. In the beginning there are many red gas particles on the right and none on the left. So, when they start to move, the chances for red to move to the left are much greater than the chances for any red to move to the right. The same goes for blue, but the other way around. In the end they will start to mix until there are as many red on the left as on the right and the same for blue. That is the situation when the chances of seeing red move towards the right are in balance with the chances of seeing them move to the left. Now we may ask whether they will spontaneously separate back into red and blue. The best guess is that they won't. The reason is that there are so many more states in which the gases are mixed than there are states in which they are completely separated, so the chances of being in the separated state are much, much lower.

Consequently, going back to the virtual particle clouds, you can see that once they start to interfere there is a much greater probability of their becoming mixed (also called entangled) than remaining separated. This process goes fast, so it takes a very short interaction time to significantly reduce the chances of getting back to the original states. This process is effectively irreversible.

12.1.1 Entanglement

Discussing the absorption of a photon brought us to the concept of entanglement. Let's see if we can make this more generic, for any kind of interaction. During

an interaction, the two wave functions of the interacting particles are said to get entangled. This entanglement can remain even when the particles fly off far away after the interaction. From the discussion above it becomes clear that the process of entanglement can be seen as the mixing of the virtual clouds of the two particles. During that mixing, they can exchange momentum leading to the particles flying off in different directions to the ones they had before the interaction (see also Fig. 10.15).

Entanglement leads to changed wave functions for both particles and, until the particles are measured, it remains unclear how the wave function has changed. However, the two particles do carry information about each other. So, when one gets measured, it will say something about the other particle, even when it has gone far away. Until the measurement we can only speak of probabilities regarding the way the wave function has evolved. The measurement leads to a definite state of one particle, and so it is said that the state of the other particle becomes definite at the same moment, which would be a “spooky action at a distance” (a term coined by Einstein) because it happens faster than light. However, there is no need for a *signal* to move through space in order to understand this [Ref. 82, 91, 92]. The change in the wave functions of the two particles gets entangled during the interaction. When one particle is measured we know what that change is and we will also know how the other wave function got changed during the interaction. We simply do not know before measuring. Note that any interaction with either particle that happens after they met is considered to be a measurement.

Entanglement is a typical quantum property. The first proof that it really exists was provided by an experiment by Alain Aspect in 1982 [Ref. 92]. However, the proof was not complete. Some loopholes remained. In 2015 the definitive proof was given in an experiment by Qutech [Ref. 91]. Today, quantum entanglement has been accepted as quantum phenomenon. It has been understood that entanglement can be used to create a computer network that cannot be tampered with without detection. Reading information on such a network is unavoidably a measurement, which will destroy the entanglement. Hence, it can always be known if the network has been breached. The first simple quantum network based on entanglement was built in 2021 [Ref. 90].

12.1.2 *Quantum Decoherence*

During an interaction between two particles, they both lose their separate states, get entangled, and fly off in a new state. The process of losing their separate states is called quantum decoherence [Ref. 52, 80] and was actually measured for the first time in 1996 by Serge Haroche et al. [Ref. 52, 55, 80].

So far, we have been talking about two particles getting entangled. However, in reality it is extremely hard to isolate a particle or a single interaction between two particles. In the atmosphere on earth, any particle has billions of interactions per second with air molecules, photons, cosmic radiation, neutrinos from the sun, etc. Even far out in the high vacuum of space, there are thermal photons from the

2.7 K background radiation, photons from stars, and so on, and the vacuum is never completely void.

The result of such interactions with the environment can be that the wave function deteriorates quite quickly. What does that mean? In a wave function, different states (or paths) of a particle have a definite phase relation. Just as we saw in the path integral, this leads to interference effects and a resulting path and state the particle is in. As long as this phase relation exists between the possible states and paths of the particle, the wave function is said to be coherent. When an interaction occurs and some virtual particles couple to (virtual) particles from another cloud, it will result in a disturbance of the wave function of the original particle. The phase relation between possible states of the particle will change or in some cases cease to exist. The result is that the wave function of the original particle loses coherence, and hence we have quantum decoherence.

Let's get back to just two particles and say we have a single wave function that describes them both. They only interact with each other. This wave function remains coherent as it contains the superposition of all possible states and paths of *both* particles *including* their interactions. Only when these two particles interact with an environment that did not belong to this wave function will it become decohered. The wave function that describes this greater system, though, will remain coherent. Taking this further, we can think of a single wave function that describes the whole universe. That wave function will always be coherent. Only when we try to look at a subsystem (say, a single particle or molecule) do we have to deal with an ever-present environment that we do not control. So, at the level of the subsystem, we experience decoherence.

The impact of the environment is substantial. Take for instance an isolated electron. Its wave function would spread across space at a speed on the order of 1000 km/s [Ref. 80, p. 117]. So that's pretty fast. It would mean that after one second, the electron will be spread across a region with linear dimension on the order of 1000 km. How can we ever localise such an electron? Measuring the electron's position would have to lead to a “collapse” of that super wide wave function. However, each interaction between the electron and the environment acts as a position measurement. Each such interaction takes away some information about the position at which the interaction took place, hence about the electron's position. Each such interaction decoheres the wave function and prevents it from spreading. It has been estimated that if we only take the interaction with photons from the environment into account, the decoherence time of an electron would be of the order of 10^{-16} s! [Ref. 80, p. 136]. So, in normal circumstances, the wave function of an electron would not spread across any measurable distance.

This is why it is so hard to measure interference effects between electrons in the laboratory. Using lots of shielding and cooling some researchers have been able to show interference effects between molecules as large as C₇₀ (Buckminster fullerene, or simply Bucky balls) [Ref. 80]. They were also able to show the decoherence effect and measure the time it took before interference was reduced to nothing.

This may still be somewhat abstract. So, let's explore a little further by looking at the essence of the problem: the measurement process. This process essentially

describes the transition from the quantum world (waves) to the classical world (particles and positions).

12.1.3 The Transition from Quantum Behaviour to Classical Behaviour

Let's resume our journey by getting back to the double slit experiment. In that experiment we saw that electrons would build up an interference pattern one by one on the screen behind the two slits, as long as they were undisturbed. Now we can see that it is very difficult to leave the electrons undisturbed for long enough to get that interference pattern. However, suppose we manage and the electron wave reaches the screen. What makes that wave "collapse" to the one position where an individual electron gets measured? Let's take it step by step.

First, we are no longer talking about an interaction between two particles. In order to measure where the electron hits the detector, the device has to consist of many atoms. The electron starts to mix its (widely spread) cloud of virtual particles with many of those atoms (see Fig. 12.2). The cloud around the photon has become part of a system that has *many more degrees of freedom*: there are many more (virtual) particles involved and hence many more states exist. The state in which the electron remains undetected is still a possible state of that system, but has become just one of

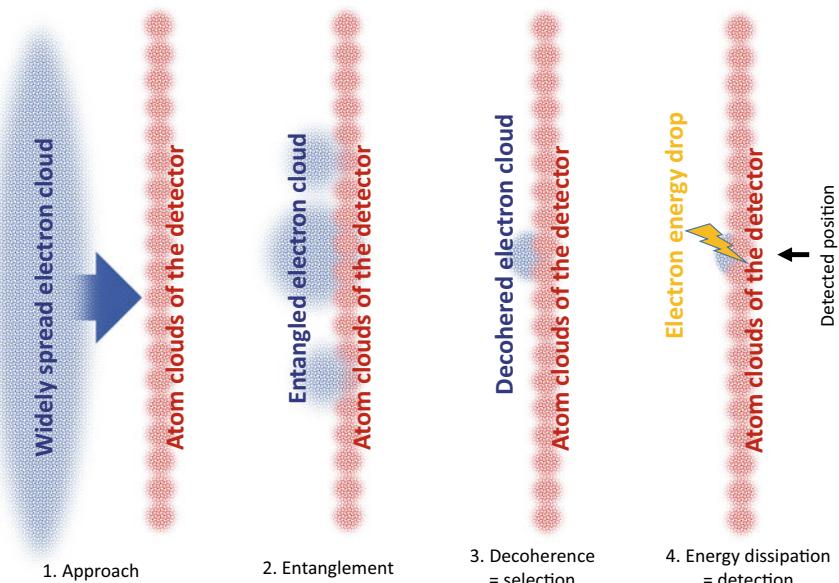


Fig. 12.2 Quantum decoherence precedes energy dissipation

very many possibilities. Among all these possibilities you would have to count those where the electron gets detected by one of the many atoms in the detector.

Step 2 is that the electron cloud gets entangled with many of the atoms in the detector. You may expect most of its energy to get deposited in a set of atoms near the spot where the electron field is strong (the top of the wave). But that would still lead to measuring a spread-out electron with most energy dissipated into the detector at the top of the wave, including some energy dissipated in other areas where the electron wave was not zero. However, that is not what we measure. We measure the energy dropped in approximately one position. Most likely that position is where the top of the wave was, but still one position.

Step 3 holds the key. While interfering with the first few atoms, the position information may still be spread out a lot, but this quickly gets reduced by the entanglement with the first measuring atoms. Remember from Sect. 4.3 that in the wave picture the change in momentum made the wave packet diversify in wavelengths and produce a narrow position. The probability that the position information narrows around the top of the electron wave is the largest, but this may also happen at other places where the electron wave is not zero. However, *each entanglement reduces the wave further*. You might see this changing the wave step by step. A changing wave is not a nice sine wave, but becomes more and more like a superposition of many waves (see also the discussion in Chap. 10). This leads to an ever clearer position. Compare this with Fig. 4.4: a position in space requires a superposition of many waves. This works the other way around too: each entanglement changes the wave, hence making the wave consist of many sines with different wavelengths. Adding up many sines reduces the uncertainty in the position. This is the way the wave becomes decohered. So decoherence can be compared to a selection process. The wave gets reduced to a single position that has been singled out by the decoherence process.

Looking at this process from a virtual particle cloud perspective, the first virtual particles get mixed with many atoms across a range of positions. The virtual particles of the atoms also mix with the electron cloud, so the net effect is still 0. At some atoms, slightly more virtual particles of the electron cloud get entangled. Where that happens, the (statistical) probability of stronger interaction increases quickly. Anywhere else, it decreases. And so, the interaction strength builds up at one place and reduces at other places. So from this point of view, the cloud also gets decohered to a single position in a statistical manner.

Only then do we get to step 4: energy dissipation. Steps 1 to 3 happen very fast, and so before any energy dissipation can take place, the wave will have been reduced to a narrow position. On an atomic scale, a long time after this game has played out, the massive interaction of the electron with nearby atoms takes place and energy starts to get transferred. That energy dissipation is picked up in the detector and so it can determine a position where the electron “hits the screen”.

Hence, the solution to the problem of the “collapse of the wave function” lies in the fact that *quantum decoherence takes place at much faster timescales than energy dissipation (see Fig. 12.2)*.

Let’s look at one more example to illustrate the difference in timescale between decoherence and energy dissipation [Ref. 80]. Take a planet in the solar system,

e.g., Saturn (my favourite). Saturn has been revolving around the sun for billions of years, hardly losing any of its orbital energy. However, the planet gets monitored continuously by photons and its position has therefore been very clear all this time. The fact that its position is always clear while hardly any energy dissipation takes place shows that it is not energy dissipation that determines the position, but rather the decoherence process. Decoherence is a purely quantum mechanical effect related to the entanglement of the clouds of virtual particles, while energy dissipation is related to the transfer of energy between already entangled systems. So decoherence always precedes energy dissipation. That's why we will never measure energy dissipation spread across the electron wave. The quantum state will have decohered before that, every time.

One finds that the level of suppression of superpositions (e.g., such as an interference pattern) increases exponentially with the number of particles in the environment that can interact with the system [Ref. 80]. This is due to the fact that these particles do not only get entangled with the system, but also with each other. Consequently, the distinguishability between states such as position states increases exponentially with time. In our example of Fig. 12.2, the many atoms are also entangled with each other. Which explains even further why the electron wave gets decohered faster when entanglement takes place with more atoms in the detector.

Let's try to generalise this concept further. In quantum mechanics we can only measure a system in its eigenstate (see Sect. 8.3). But as long as it is not measured, the system evolves as a wave, described by the wave function. The Born rule defines the probability of finding the system in one of these eigenstates when measured. So where do these eigenstates come from? Who has defined what are supposed to be the eigenstates? Now that we understand a bit about decoherence, we can start to formulate an answer to those questions [Ref. 80].

One of the most common eigenstates is the position of a particle. We saw already how the position can get monitored by the continuous presence of billions of interactions with the environment. This means that the position information does not get stored in that environment once or twice, but is abundantly available and redundant. So, we can figure out the position of a particle at any time by measuring the environment. For example, we find the position of Saturn by measuring the photons that got deflected off it. With each interaction with the environment, the particle's wave function decoheres to the position where the interaction took place. Afterwards it can evolve, but the next interaction will already have taken place. So, position is a robust state: it is there to stay due to the many interactions that are forever reducing its wave function to an actual position. Superpositions of position (like an interference pattern spread across positions) are not robust since they are challenged with every interaction taking place. This makes the (classical) position of a particle a robust state that is redundantly available in the environment to measure. And that, in turn, makes it an eigenstate. Eigenstates are those states that are most redundantly stored in the environment. That makes them the most measurable states in the environment. Applying the Born rule, we can now define the probability of finding a particle at a certain position. This is made meaningful by the continuous probing by the environment.

The measuring device is just as much part of the environment. It is designed either to measure the redundant information in the environment or to enlarge the measurement of a single particle. The enlargement process is similar to monitoring by many particles in the environment. In both cases, many states of the environment or the device average out. The more averaging, the clearer the measurement and the clearer the single state (e.g., position) that comes out. Our senses are such devices. Compare this to the pressure of a gas: it is sharply defined and measurable, while in fact it is a statistical average of the motions of billions of gas molecules. Finally, the device is designed to measure A or B and therefore has many states for A and many for B and only a few for other states. So statistically, we arrive at A or B, a single outcome rather than a superposition.

Another example are the energy levels of an electron in an atom. The electron is heavily entangled with the nucleus and with the other electrons in the atom. It gets continuously probed through that entanglement. However, it does not get probed for its position. The information exchanged in that entanglement does not reveal its position, but it does reveal its energy level. So, the electron gets probed for its energy level. Why does the entanglement within an atom show a different behaviour compared to measuring the position of an electron in a detector? To measure a position requires a change in the momentum. Only a change in the momentum increases the uncertainty in the momentum, thereby decreasing the uncertainty in the position. An electron captured within an atom does not change its momentum within the atom. This is because the electron can only exist within the atom in a state of definite wavelength (see Sect. 13.1), hence of definite momentum. The momentum of the atom as a whole can change, but not that of the individual electron captured in an “orbit” within the atom. Clarity about its momentum translates into *a lack of clarity* about when the electron is in a specific place. Uncertainty about time increases the certainty about the energy level. In general, we have to consider position and time as a four-vector that has an uncertainty relation with the momentum—energy four-vector. So, the wave picture explains why sometimes energy—momentum gets defined by entanglement and sometimes time—position. Generally speaking, free particles can continuously change energy and momentum during interactions. Hence, their position in space–time is an eigenstate. Particles that are bound in a potential well must have certain wavelengths and frequencies (see also Sect. 7.1). Hence, their momentum and energy are eigenstates and their position in space–time is unclear. Therefore, this entanglement of an electron within an atom can distinguish between energy levels.

Consequently, superpositions of different energy levels will not last, but superpositions of position do last in the atom. This is why the electron stays in such a well-defined energy level, while its position in the atom is never clear and can only be summarized as a probability of finding the electron somewhere in the atom. Consequently, the energy levels are the (robust) eigenstates of the atom. Of course, the atom as a whole gets abundantly probed for its position by the environment. But this does not generally distinguish (or reveal) the position of each electron within the atom itself.

Since eigenstates emerge from environmental probing of the particle wave, the process by which eigenstates are created is called “environment-induced superselection” [Ref. 80]. So, it’s the interaction with the environment that selects the robust states that become the eigenstates. The interaction is a process of entanglement leading to decoherence. So, we can say that the cloud of virtual particles around a particle is responsible for entanglement with the environment, changing its own wave function, and it is from this that an eigenstate emerges.

Since the eigenstates are actually classical observables (such as position or energy level), this clarifies the transition from quantum behaviour (interference) to classical (robust eigenstates). For a long time, scientists could not explain the transition from quantum to classical. Measuring a wave function would lead to a “collapse of the wave function”, but the collapse process was unknown. Entanglement and quantum decoherence have given us some insight into the way this process takes place. Quantum behaviour is all around us. Particles are actually just clouds of waves and disturbances. But their classical appearance is caused by the way the virtual cloud of a particle gets entangled with other particles. On the macroscopic level a lot of particles play a role in that process. Such a large number of particles quickly decohere the quantum state of any system we want to measure (or see). In the end only the classical information actually emerges from it.

12.2 Tunnelling and Decoherence

In Sect. 7.1 we covered the concept of tunnelling. We made clear that when a quantum is made of a variety of waves with different frequencies, there can be some waves that have enough energy to pass a potential barrier. In particular, for a short period of time, a quantum is made up of a variety of different waves (see Chap. 4). The shorter the time interval, the greater the range of frequencies (hence energy). However, this process does not explain why only some of the waves making it to the other side can make the *whole* quantum move to the other side of the potential barrier.

So, let’s revisit tunnelling from the perspective of virtual particle clouds and decoherence to see if we can answer that question. To do so, we will make use of the specific feature of virtual particles (or disturbances) that they exist off-shell (see Chap. 10). This fact has a deep implication. Off-shell, the relation between energy, mass, and momentum is no longer strictly obeyed. What does that mean?

First of all, it means that a virtual particle may have any energy. Hence, it can cross a potential barrier that the real quantum could not cross. So, when a quantum actually exists of a cloud of virtual particles (disturbances), there is a probability that it might move all its properties across the barrier one by one, since the cloud’s individual members are *all* off-shell. It is only added up that they act as a single on-shell particle.

Next, what will happen if some of the virtual particles on the other side of the potential barrier get entangled with clouds of other quanta on that side (see Fig. 12.3)?

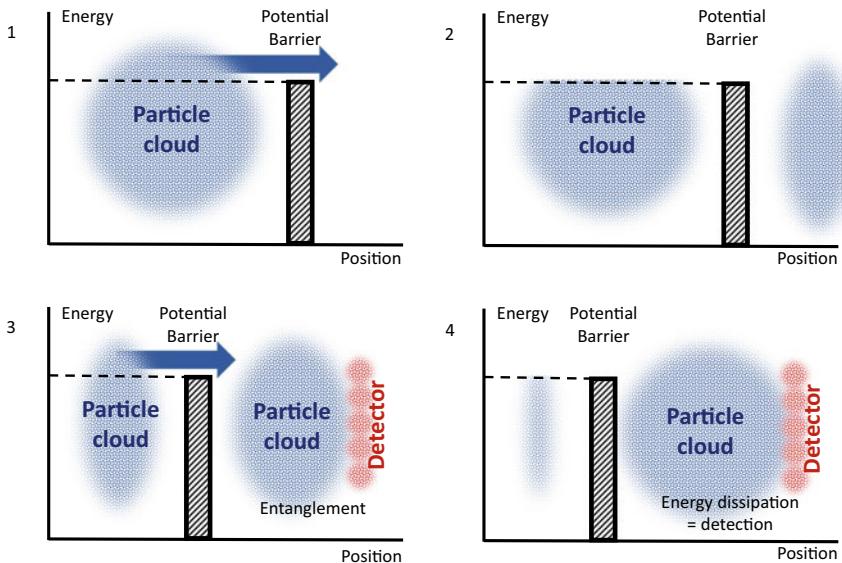


Fig. 12.3 Tunnelling. In step 1, those virtual particles with enough energy cross the potential barrier. In step 2, part of the virtual cloud is on the other side and only lower energy virtual particles remain on the left. In step 3, the particle cloud on the left spreads its energy across the cloud and again some have enough energy to cross to the other side. When there is a detector on the other side, entanglement starts to take place. In step 4, most of the cloud has crossed and entanglement has ensured that the probability of finding the particle on the left-hand side quickly deteriorates, as most entangled states are on the right-hand side. This, in turn leads to energy dissipation in the detector and thus detection of the particle on the right

For example, suppose we have a measuring device on the other side. The virtual particles of the quantum that crossed the barrier start to get entangled with the measuring device. This leads to decoherence of the quantum. Over a short period of time the number of states in which the quantum has interfered with the measurement device becomes far larger than the number of states in which the quantum remains on its original side of the barrier. So, one by one, the virtual particles cross the barrier. Finally, the energy of the quantum gets dissipated in the detector and the quantum gets measured on the other side of the barrier.

However, not many virtual particles will cross the barrier initially. So, the probability for this event to happen really depends on how far off-shell the virtual particles need to be before they can cross. The further off-shell, the less likely the process becomes.

There is another aspect to being off-shell that has implications for the timescale of tunnelling. Basically, on mass-shell, a particle behaves in the way we are used to: a particle's velocity remains below lightspeed. Compare this with the blue line in Figs. 6.5 and 10.3. It will not cross the red line that describes the situation for massless particles such as photons. In Sect. 6.2 it was explained that being on-shell means having a speed below c .

However, on the right-hand side of the red line, the particle would move faster than light! So, does this mean that virtual particles can move faster than light because they are off-shell? Well, there is a certain probability for virtual particles to do so. However, the more c is exceeded, hence the more off-shell the virtual particle gets, the more that probability deteriorates [Ref. 9, Chap. I.3]. Potentially, this means that tunnelling could take place faster than light. In practice, experiments show that the tunnelling time can be less than the barrier thickness divided by the speed of light, without violating causality [Ref. 81]. So, we may take this as experimental verification that virtual particles can move faster than light.

So, it seems that decoherence has an influence on tunnelling. But can that really be true? Suppose there is no measurement device and no environment that can be probed for the existence of the quantum on the other side of the barrier. Then we cannot say whether the quantum has tunneled or not. However, in theory we can say that a wave function can tunnel from A to B and back again. Therefore, it could exist on both sides. This implies that there is a probability of finding the quantum on side A as well as a probability of finding it on side B. Actually, finding it implies measuring it and hence a decoherence effect in the act of measuring. So, it is not really possible to assess the influence of decoherence on tunnelling. The only thing decoherence can be accused of is making it clear on which side of the barrier the quantum resides.

However, not so fast! There can be different types of decoherence. One type that is hypothesized is called intrinsic decoherence. This is a type of decoherence that is not environmentally induced. It would imply that wave functions can decohere on their own. Consequently, there would actually be some “collapse of the wave function”. Such a property of the quantum world would significantly change the views presented in this book. If such intrinsic decoherence does exist, it is expected that it would actually influence the tunnelling process [Ref. 89]. It could increase the tunnelling rate, depending on the intrinsic decoherence time. The shorter this time, the higher the probability of tunnelling (up to a theoretical limit). Maybe this can be understood by the decoherence process actually forcing the quantum to be on one or the other side. The reason why this is investigated is that the measurement of tunnelling times can give a clue about whether or not intrinsic decoherence exists [Ref. 89]. That might give us conclusive information about whether environmental decoherence is solely responsible for the quantum to classical transition. Experiments so far have indicated that environmental decoherence is in any case an important and dominant process that explains the classical behaviour we usually encounter in the macro world. But is it the only process that is responsible for this transition? That is at present a key question.

12.3 Quantum Computing

A quantum computer distinguishes itself from a classical computer by using entanglement and superposition to make calculations. So, let's see how that works [Ref. 84].

A classical computer uses 0's and 1's to represent information. For example, 011 can represent the number 3. Adding $3 + 3$ would result in $011 + 011 = 110 = 6$. Each 0 or 1 is called a "bit". By analogy, a quantum computer uses so-called "qubits". These are not the same as bits. Where each bit in a classical computer is either 1 or 0, a qubit is more or less both at the same time, as well as any combination of them. We can clarify this by looking at the way the 0's and 1's are represented in a physical device. In a classical computer, a 0 or 1 is represented by the position of one switch in a computer chip. The switch is either open (0) or closed (1). And so, a single switch can represent one classical bit. A qubit is not represented by a switch but by the state of a quantum system. For instance, a state of electron spin. This state can be up (0) or down (1) when we measure it (it's eigenstates). But as long as we do not measure its state, all combinations are possible. Each combination represents a probability between 0 and 100% of measuring the system to be in 1 or 0. The state of such a system can be described by

$$|\text{state}\rangle = a|0\rangle + b|1\rangle$$

where a and b represent the amplitudes for finding the system in $|0\rangle$ resp. $|1\rangle$. Graphically, we can visualise all the possible combinations on a sphere, called the Bloch sphere (see Fig. 12.4). In this visualisation, we map a to the angle around the z-axis ($0\text{--}360^\circ$) and b to the angle around the x-axis (-90° to $+90^\circ$). That way we get a complete sphere that represents all possible states for this system.

This is still somewhat abstract. Let's see if we can make it more visual. Both the states spin up (= 0) and spin down (= 1) can be visualised as the electron wave turning around an axis (see Chap. 13). For the sake of simplicity, we forget here that an electron spins on a Möbius strip. Then how should we see the combined state? Let's depict both states as water waves (Fig. 12.5). When these waves meet, they create the well-known interference pattern. What you see is the superposition of the two waves. Let's go one step further. What if we put a stick in the water? Clearly, both waves (and their superposition) are influenced by the stick at the same time. If we now compare the stick to a computer operation or algorithm, it becomes clear how a computer operation can act on two states (and their superposition) at the same time.

So, let's take this concept and work out how a one-qubit computer might work. A single qubit can hold two numbers 0 and 1 in superposition at the same time. So when we apply an operation to that qubit, we operate on two numbers at the same time (see Fig. 12.6). Such an operation might be "select the largest number". *Before* the operation, the qubit would equally likely produce 0 as 1 if we measured

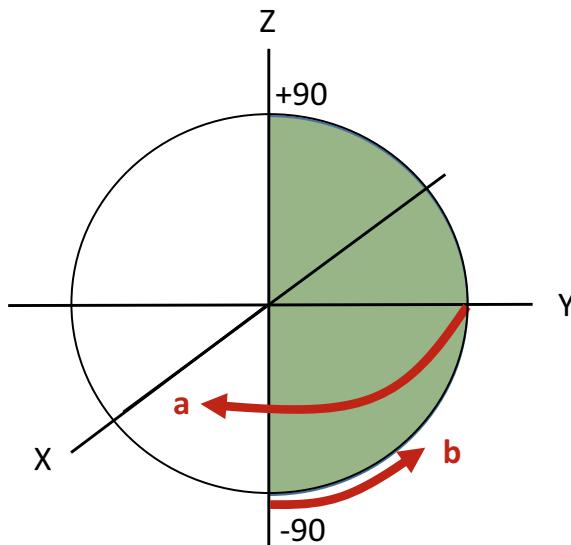


Fig. 12.4 The Bloch sphere

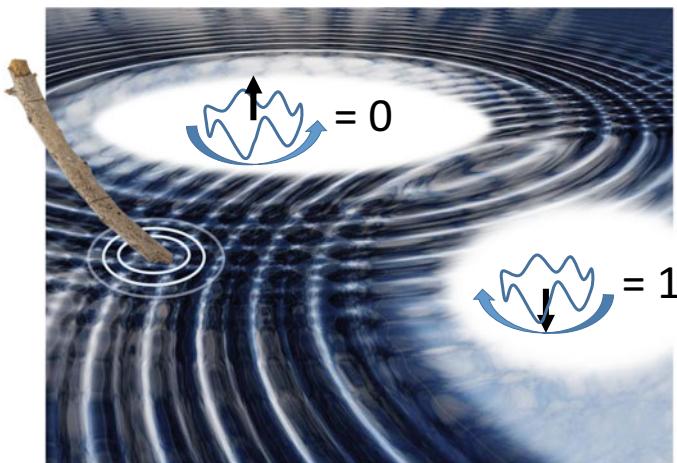


Fig. 12.5 Representation of the two possible electron spin states up and down by two waves on water. The total quantum state of the electron (when not measured) consists of the superposition of the two states. A stick in the water represents an operation (or algorithm of operations) on the two states (which is not a measurement as it does not provide information about the spin state of the electron). That operation clearly influences both states and their superposition

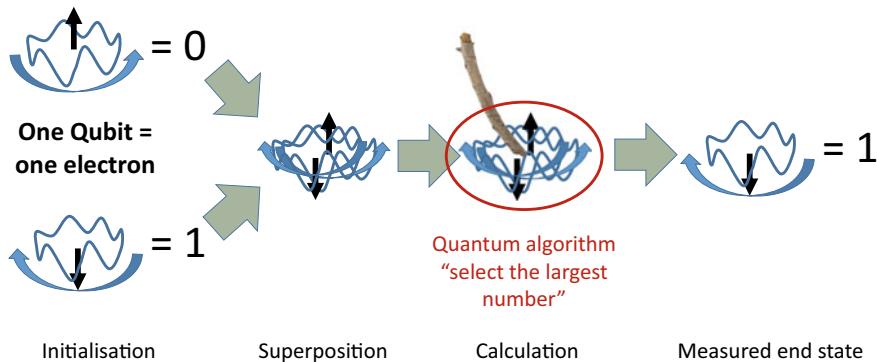


Fig. 12.6 A one-qubit computer can operate on two numbers at the same time using a quantum algorithm

it. The quantum operation, however, changes the odds. After the operation we would measure 1 in 100% of the cases.

All this is not yet very spectacular. What if we were to build a 2-qubit computer? In order to do the same thing with two qubits, we first need to get the qubits entangled (see Fig. 12.7). Getting the two qubits entangled means that their respective wave functions become one combined wave function for as long as they do not get measured. While they are entangled, we can apply quantum operations to them. These operations are not the same as measurements! The operation performed will never reveal the state of (one of) the qubits. As long as the state is not revealed, the qubits can remain entangled during the operation.

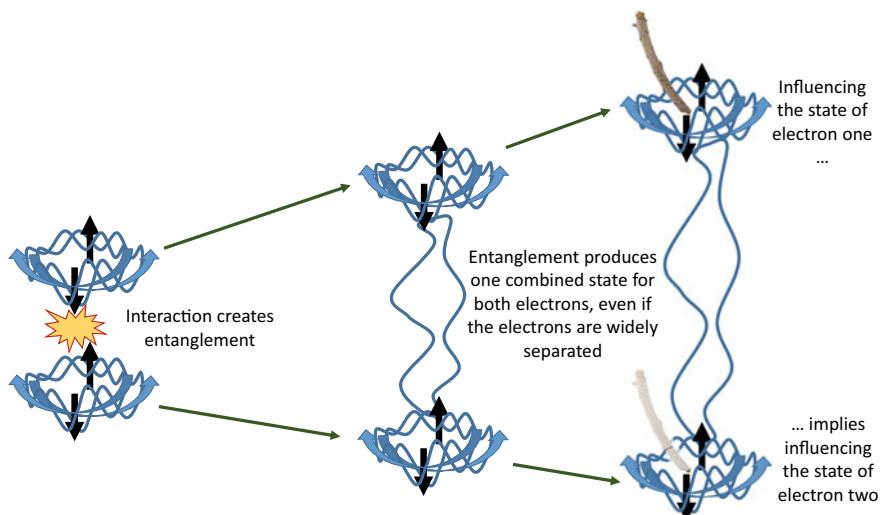


Fig. 12.7 Getting two qubits entangled

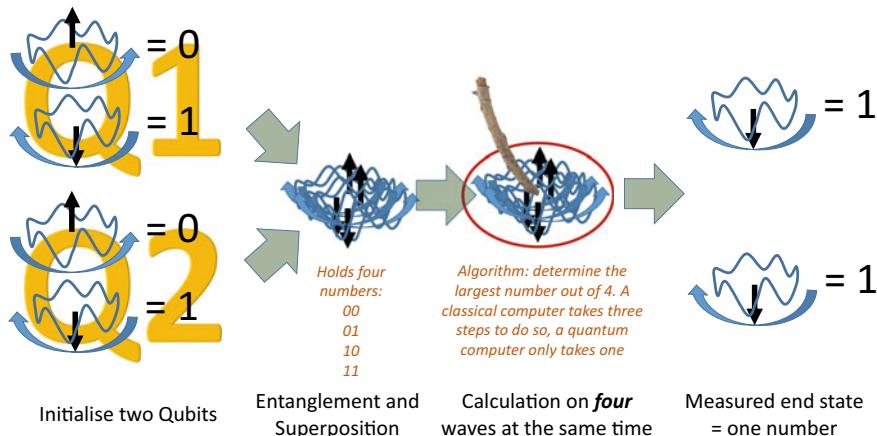


Fig. 12.8 A two-qubit computer

Now let's see how a two-qubit computer works (see Fig. 12.8). We first need to initialise the qubits. This means we may (or may not) bring them into a certain state. Then we need to create the entanglement between the qubits. Once that is done, the entangled 2-qubit state can represent four numbers at the same time. Next, we can apply our algorithm to select the largest number. After this operation we can measure the end state and we will find the largest number amongst the four. What is important is that it takes only one operation to make this selection. A classical computer needs three comparisons to find the largest number. In essence this is because a classical computer can only ever compare two numbers at a time.

I can imagine this still does not impress you. In order to do that, we need to beef up our number of entangled qubits. Three qubits would be able to hold 8 numbers (2^3). With 8 qubits we would have 256 numbers. 50 qubits give us 1.13×10^{15} numbers. Now try to find the largest prime number amongst these. A supercomputer could still do this in approximately the same order of time as our quantum computer. But if we had 100 qubits, a modern supercomputer could take up to 40 million years to make a comparable selection. With 200 qubits, the supercomputer would take longer than the age of the universe.

I hope by now you have noted one very important thing. In our examples we only talk about a selection, not about parallel computing per se. A quantum computer may be able to operate on a huge number of inputs at the same time, but it will always produce just one answer! The answer you get is the measured end state of all the qubits. This is just one number. So, a quantum computer will not be able to perform many calculations in one go, since it would have to produce the same number of answers in the measured end state, and this it cannot do.

This means that a quantum computer is primarily able to perform a specific type of task very efficiently. These tasks fall into the category of problems we could call “selection problems”. As in our example, finding the largest number amongst a set of numbers is a selection problem. There are many such problems and many of these

cannot be solved with classical computers. One example is the classical “travelling salesman problem”. In this problem, the salesman needs to visit a lot of customers throughout the country and wants to find the most efficient route to do so. One could formulate the problem as follows: find the shortest route that visits every customer’s location exactly once and ends at the starting position. This can take a classical computer a surprising amount of time since in principle it has to consider every option (although there are handy algorithms that considerably shorten the time). A quantum computer could consider all possible routes in one go and solve the problem in a very short time. The outcome would be one result: the shortest route. Finding a shortest route or most efficient configuration has many applications in almost every subject.

Another type of problem for which quantum computers are particularly useful are quantum simulations. One tries to map the quantum states of the energy levels of a complex quantum system (such as a molecule) to the initial states of the quantum computer and let it evolve. The end result of the process in the quantum computer will very likely represent the state of lowest energy of the system. For example, it should show how a molecule will fold into its stable lowest energy state.

Here are some examples for which we could use a quantum computer:

- Security: prime factorization—what are the two largest prime numbers that multiply together to make a given number? This would break a particular kind of encryption (RSA encryption). Not all kinds of encryption can be broken with a quantum computer, however.
- Chemical (quantum) simulation: what is the optimal configuration of a molecule? This has been solved at present only for the simplest of molecules using classical computers.
- Finance: what is the investment scenario with the lowest risk (taking a lot of factors in consideration)?
- Artificial intelligence: what is the optimal configuration of a neural network?
- Chips: what is the most optimal architecture of a microchip?
- And many others.

So far, we have not discussed the role of quantum decoherence. And it turns out to be decoherence that makes it so extremely difficult to build a quantum computer [Ref. 83]. Why would that be? Basically, our quantum computer only works when we manage to get all qubits entangled and behave as one wave function. Moreover, it has to do this while applying the entire algorithm until the final stage: measuring the end state. So, the more complicated the algorithm, the longer the entangled state needs to persist.

When we discussed quantum decoherence, it became clear that there are always many interactions with the environment that would decohere the wave. In fact, decoherence timescales appear to be extremely short. So, to get a quantum computer functioning, we would need to shield it from the environment extremely well. This includes all types of radiation, contact with other molecules, external electromagnetic fields, heat, etc. So, a quantum computer needs to be extremely cooled (down

to 10 milli kelvin), placed in a very good vacuum, and shielded from electromagnetic fields. And even then, the stability of present-day quantum computers is low. To some extent the stability problem is circumvented by repeating the calculation 200—1000 times. The answer that comes up >95% of the time is probably the right answer. Researchers building quantum computers consider the stability problem as more important than the number of qubits. Fewer qubits with higher stability are often more valuable than more qubits with lower stability.

12.4 Schrödinger's Cat

As a summary, let's return to Schrödinger's poor cat. The discussion in this chapter, including the discussion of the quantum computer, illustrates how predominant quantum decoherence actually is. To maintain a quantum state is so hard that it shows why we constantly perceive a classical world around us. Quantum behaviour is rare due to decoherence. It also shows that excluding the environment (e.g., in order to perform a “pure” experiment on a well-prepared sample) is not as easy as it seems. The effect of the environment and its many interactions is something to be aware of. The thought experiment about Schrödinger's cat basically did not take that environment and the effects of (mutual) interactions into account. When one includes the environment and the many interactions in large systems, one can better solve the paradox that was put forward. In the case of Schrödinger's cat, it is not the person outside the box who determines whether the cat lives or not. This is determined inside the box by the many interactions between the measuring device, the cat, and the killing device. Even on a microscopic level, (part of) the cat is not in a state of superposition between alive or dead. Not even the mechanism that produces the deadly gas is in a superposition of two states. Probably, only the radioactive atom that should trigger the deadly switch could be in a superposition, but due to the many interactions with the environment, even that will not long be the case. Hence, quantum decoherence reduces the question of Schrödinger's cat to a purely classical problem.

Chapter 13

Spin Makes Up Bosons and Fermions



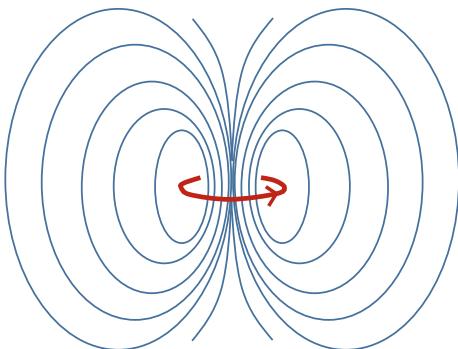
So far, we have been ignoring particle characteristics such as spin, helicity, and chirality. We will view these characteristics from a wave perspective instead of the more usual particle perspective that prevails in quantum mechanics.

13.1 What Is Spin?

It is tempting to view “spin” as the spinning of a particle. Often, for simplicity’s sake, it is treated as such. In more serious physics books spin is called an internal degree of freedom. This is the most general way to describe it. This is done because we really do not know exactly what spin is. We know from experiment and theory what the characteristics of spin are and from a mathematical point of view, treating it as just another degree of freedom works well. Some of those characteristics are:

- Spin is quantized, so not each “spinning speed” is allowed.
- Particles have a definite spin that dictates what spin states they can assume, for instance spin $\frac{1}{2}$ dictates that the particle can be in the spin states $+\frac{1}{2}$ and $-\frac{1}{2}$.
- A charged particle with spin has a magnetic field proportional to its spin. A charge that runs in a circle produces a magnetic field (see Fig. 13.1). The speed with which it goes around determines the strength of the magnetic field. This comparison leads us to the picture of a spinning charged particle that produces a magnetic field.
- Particles with spin $1/2, 3/2, 5/2$, etc. are called fermions.
- Particles with spin $0, 1, 2$, etc. are called bosons.
- Fermions and bosons have clearly distinguishing characteristics dictated by their spin.

Fig. 13.1 A charge that goes around in a circle (red) produces a magnetic field (blue). So, it is tempting to explain the fact that a charged particle such as an electron produces a magnetic field by viewing it as a spinning particle



So, it is clear why a particle view is tempting. In particular, it is the idea of the spinning particle as a dynamo creating a magnetic field that suggests that we may view spin as a feature of a revolving particle. However, at the scale of fundamental particles such as electrons and quarks, we do not know how to view such a particle. In the previous sections we have seen some strange behaviour that cannot be easily combined with this picture. And then there is the quantization of spin. Why is that? The spinning particle idea does not give us an answer to that. Another question is: why do fermions behave so differently from bosons? The idea of a revolving particle does not help us there either.

So, let's look at a wave approach to understanding spin. This too is just a picture that helps us understand what is going on, but it is by no means necessarily the truth of what spin is. Keep in mind that we can only measure the characteristics, and we cannot see directly what spin is.

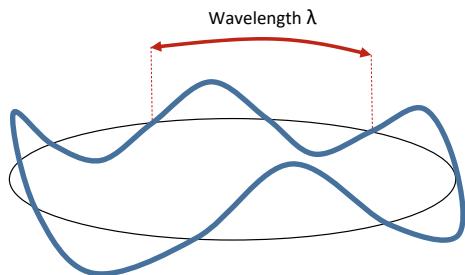
13.1.1 Orbital Momentum in the Atom

As an example, we will first treat the orbital momentum of electrons in an atom. Orbital momentum is not the same as spin momentum. Orbital momentum concerns an electron wave going around the atom, while spin momentum is an internal degree of freedom of the electron itself.

If we quantize *orbital* momentum, we can view a particle such as an electron going around a circle in the atom. Viewing the electron as a wave, such a wave can only go around a circle that fits the wave exactly a whole number of times. If it does not, the wave will eventually (after going around a number of times) interfere negatively with itself and fade out. So, for an electron wave, the only stable way to exist in orbit is to choose those orbits that fit exactly one or more whole waves (see Fig. 13.2).

The result is that electrons can occupy only certain orbits at a well-defined distance from the centre of the atom. The smallest orbit the electron can occupy is the one that harbours only one wavelength of the electron wave. There is no way the electron

Fig. 13.2 In this example four wavelengths ($4 \times \lambda$) fit on a circle



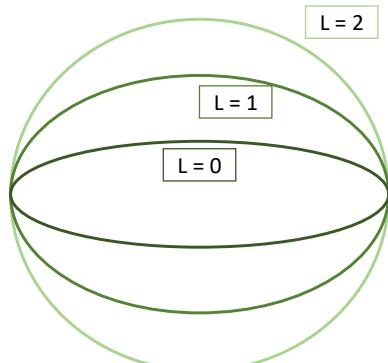
can occupy a smaller orbit, since it can no longer fit in a whole wavelength. The next (bigger) orbit is the one that can contain exactly two wavelengths.

Each orbit as discussed so far is indicated with a quantum number n . The potential energy of each orbit is proportional to $1/n^2$ so that we see quantized energy levels appear. It takes more energy to free an electron from the smallest orbit than from a larger orbit. This is logical, since one has to overcome a larger potential difference to free the electron from a smaller orbit. So, $n = 1$ agrees with the smallest orbit to fit the electron wave on. The next orbit is $n = 2$ and has an energy $2^2 = 4$ times lower.

However, this is not the whole story. Orbitals can be degenerate, meaning that there can be more states that occupy each energy level. How is that possible? It turns out that there is a whole class of orbits that have the same energy as a circle. These are elliptical orbits whose long axis equals the axis of the circle (see Fig. 13.3). This means that, e.g., a planet orbiting in each of these ellipses around the sun would have the same energy. The difference is that the sun would be placed in the centre for the circle, but not for the ellipses. Another difference is that these ellipses do *not* have the same orbital momentum. So even though, each mass in each ellipse would have the same energy, they would differ in their orbital momentum.

Let's see how that works for electron waves. First of all, the case of $n = 1$. In this case, only one wavelength fits the circle. We can find all sorts of ellipses that have the same energy. However, none of these would fit a wavelength since they are all much smaller. So, $n = 1$ just represents one state and is not degenerate, meaning that no other states are possible at this energy level.

Fig. 13.3 All these ellipses have the same long axis. The long axis determines the energy of the ellipse, so each of these ellipses have the same energy. When a mass orbits like this around, e.g., the sun, the total energy (= potential energy + kinetic energy of the mass) is the same. However, each of these ellipses would represent a different orbital momentum



So, take the case $n = 2$. We can now fit 2 wavelengths on a circle. This circle is bigger, so it is further away from the nucleus. Hence, the energy needed to free an electron from this orbit is less and so we say that the electron is on a higher energy level. We can again find many ellipses that are smaller in circumference but represent the same energy.

None of these will fit two wavelengths, but we can also try to fit one wavelength on them! If we do, at first it seems as though we cannot. After all, the circle has a circumference of $2\pi r$, fitting 2 waves, so one wave would have a length $\pi r = 3.14\dots \times r$. The smallest ellipse we can think of is a line through the centre with a length equal to the diameter of the circle. The diameter is $2r$, and the wave would have to go back and forth along the diameter, which is $2r + 2r = 4r$. So, one wave would not even fit the smallest ellipse we can think of. So, is this energy level not degenerate?

Wait a minute! The smaller ellipses also have a smaller orbital momentum. And we have seen that smaller momenta correspond to longer wavelengths ($P \sim 1/\lambda$)! Consequently, when we consider smaller ellipses, we have to take into account the fact that the electron wavelength will be longer!

Since its energy will be the same (and hence its frequency F will be the same), the “velocity” v of the wave would have to increase to have a longer wavelength λ , according to the formula $F = v/\lambda$. Compare this to the solar system. It is known from elliptic orbits that when a comet gets closer to the sun, its velocity increases. That is the whole idea behind orbital momentum: it is defined as the distance to the sun \times the velocity of the comet. Hence, because orbital momentum is conserved, the closer it gets, the faster it goes. On the other hand, the farther away it is, the slower it goes. On average, the orbital momentum is lower for an elliptical orbit than for a circular orbit. So, at shorter distance, the higher speed does not weigh up against the shorter distance, and the distance \times speed is smaller than for the circular orbit.

So, in the end we can find an ellipse that will fit one wavelength, with a lower orbital momentum, but the same energy. If we label the circle version of this orbit and the ellipse version each with a quantum number l , we get $n = 2$ having two energy levels $l = 1$ (circle) and $l = 0$ (ellipse).

But we are not there yet! There are more possibilities with the same energy. Take a look at the circle for $n = 2$. It contains 2 wavelengths. Now we have to start thinking in three dimensions! There is not just a circle around the nucleus that has the same energy; there is a whole sphere with the same distance to the nucleus that corresponds to the same energy. On that sphere we can find a smaller circle that contains one wavelength. There are two of those: one on the “northern hemisphere” and one on the “southern hemisphere” (see Fig. 13.4). So, we find three circles with the same distance to the nucleus and therefore the same energy. These circles get the quantum number M . Since these are all *circles* with the same energy, they also have the same orbital momentum.

So, we end up with the $n = 2$ level being four times degenerate: a circle fitting two wavelengths that is accompanied by two smaller circles fitting one wavelength with the same energy and orbital momentum. And we have an ellipse fitting one wavelength with the same energy but smaller orbital momentum. All this is summarized in Table 13.1.

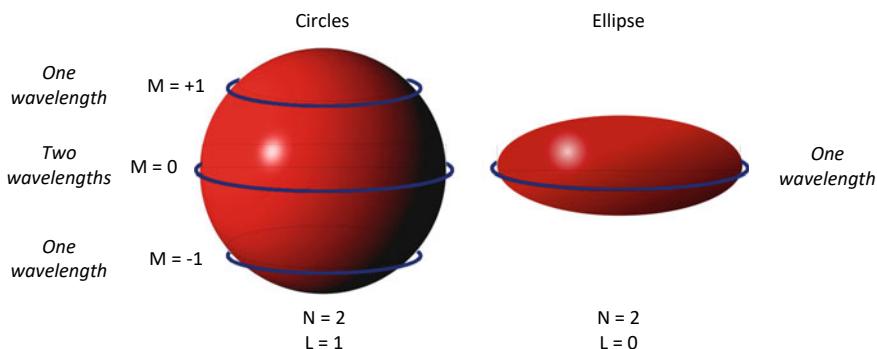


Fig. 13.4 All possible states with the same energy at $n = 2$

Table 13.1 Circles and ellipse with energy level $N = 2$, but different orbital momenta and types of orbits

Type of circle	Energy level	Orbital momentum	Types of orbits	No. of wavelengths
Ellipse	$N = 2$	$L = 0$		1
Top circle	$N = 2$	$L = 1$	$M = +1$	1
Middle circle	$N = 2$	$L = 1$	$M = 0$	2
Bottom circle	$N = 2$	$L = 1$	$M = -1$	1

This system continues. When we go to $N = 3$, we find a circle that fits three wavelengths. This circle is on a sphere with a smaller circle in the northern hemisphere that fits two wavelengths and an even smaller one that fits one wavelength. The same goes for the southern hemisphere. Then there is an ellipse that fits two wavelengths. But this ellipse can also be seen as an ellipsoid (just like the sphere). And on this ellipsoid, we can find a smaller ellipse that contains just one wavelength. Finally, we can find an ellipse that is more squashed than the previous ellipse. We say that this ellipse has a higher eccentricity, meaning that it looks flatter. Put differently, the short axis is shorter compared to the long axis. Since this ellipse is more squashed, it has an even lower orbital momentum and thus longer wavelength, and it will fit one wavelength at the same energy. This is summarized in Table 13.2.

This system continues. In general, we have that, for each N there are N possible values of L , and for each L there are $2L + 1$ possible values of M . When we add that up, we get a total of N^2 states that have the same energy. Each state can hold 2 electrons (as we will see later). Consequently, we get the so-called atomic shells in which we can find a different number of electrons in each energy level (see Table 13.3).

The atomic shells and the way they are filled by electrons create the chemical properties of the atoms. All these properties are a direct result of fitting electron waves on circles and ellipses that have the same energy.

Keep in mind what these electron waves really mean: they are waves of the electron field. The amplitude of the wave at each point around the circle indicates the field

Table 13.2 Circles and ellipses with energy level $N = 3$ and their types of orbit and momenta

N	L	M	Type of circle	No. of wavelengths
3	2	+2	Smallest circle, northern hemisphere	1
3	2	+1	Circle, northern hemisphere	2
3	2	0	Central (largest) circle	3
3	2	-1	Circle, southern hemisphere	2
3	2	-2	Smallest circle, southern hemisphere	1
3	1	+1	Smallest ellipse, northern hemisphere	1
3	1	0	Central (largest) ellipse	2
3	1	-1	Smallest ellipse, southern hemisphere	1
3	0	0	Ellipse with higher eccentricity	1

Table 13.3 Overview of the number of orbital momentum states each energy level can hold, and consequently, the number of electrons that can maximally be found in each energy level

Energy level	Contains	That can hold
$N = 1$	1 state	2 electrons
$N = 2$	4 states	8 electrons
$N = 3$	9 states	18 electrons
$N = 4$	16 states	32 electrons
Etc.		

strength. So, when the wave is at a peak, the field strength is high. The field strength in turn determines the level of interaction with other fields. Hence, it impacts the chances of interacting. High field strength means a high probability of interacting. A high probability of interacting translates into a high probability of “finding the particle at that spot”. In quantum mechanics the wave function is interpreted as the probability of finding a particle there. So effectively, the waves present us with a probability distribution of where the electron “is” in the atom.

You may wonder why the circles on the northern or southern hemisphere could be valid “orbits”. Classically, a planet could not orbit like that around the sun. The fact that we can fit a wave on these “orbits” does not explain why an electron could actually be on one side of the atom all the time. The point is that since all these states have the same energy (they are degenerate), they are indistinguishable. We have seen that indistinguishable states in quantum mechanics will mix up. So, an electron could be in any of those states at one point in time and in another state at another moment. This means that the electron does not stick to one side of the atom.

Only when we separate these states could the electron remain in such a state. The electron waves that are moving in orbits are essentially electric currents moving in an orbit. Such a current produces a magnetic field. When we apply a magnetic field from outside, it will interact with the electron wave magnetic field and separate the states with different values for M. This means that the interaction with the magnetic field causes the energy level of each state to be different. This is how we can separate

the different states. Consequently, the electron will remain in one state. It can also remain in that state, since it will be kept there by the magnetic field applied from outside.

The model explained here is based on the so-called “Laplace–Runge–Lenz vector” (LRL vector). This is the vector that is associated with an ellipse (of which the circle is a special case). This vector is a constant of motion, meaning that it has the same value everywhere on an elliptic orbit under certain circumstances, viz., the force must be a central $1/r^2$ force. Such orbits are called Kepler orbits, after the astronomer Johannes Kepler, who lived in the seventeenth century. He worked out such orbits for planetary motion.

The LRL vector was used in the first derivation of the spectrum of the hydrogen atom. It was based on quantum mechanics, but before the Schrödinger equation had been found [Ref. 69]. The results of this model were in excellent agreement with experiment. However, the hydrogen atom is simple. It has only one proton in the centre and one electron moving around. This is the ideal force one can apply the LRL vector to. In larger atoms, the nucleus consists of more protons so the electromagnetic force is created by more than one charge. Moreover, there are more electrons around, and these influence each other. We could say that each electron creates an electromagnetic field that affects the other electrons, thereby lifting the degeneracy of the energy levels. In other words, there are small differences in energy between the states with different values for L and M . Later on, quantum mechanics and perturbation theory were used to describe more complicated systems, and the method based on the LRL vector is no longer used very often. Its principle is still valid though and it gives great insight into how everything works. It is a good basis for explaining atomic structure using the wavelike nature of electron fields.

13.1.2 *The Origin of Spin*

Orbital momentum has a clear origin. We could fit the electron wave on different orbits around the nucleus. What, then, does it mean when a particle spins by itself? After all, the particle is really a wave in a field. And how would a wave be spinning? To understand how this can work, we first have to go back to special relativity.

In relativity there is a difference in the way velocities add up. When we ride on a train and we throw a ball we can simply add up the velocities of the train and ball in ordinary space. But close to the speed of light, velocities add up differently. The change from one frame of reference to another (like jumping onto a train with a different velocity) is known formally as a “boost”. In a boost, you give an object a different speed. It is not the process of acceleration that is described here, but the effect of having another speed. In this context, a Lorentz transformation is closely related to boosts. A Lorentz transformation describes how the space and time coordinates of one frame of reference must be “transformed” to the space and time coordinates of another frame of reference (moving at a different speed). So, it shows how to relate the coordinates of two frames when one of them is “boosted” relative to the other.

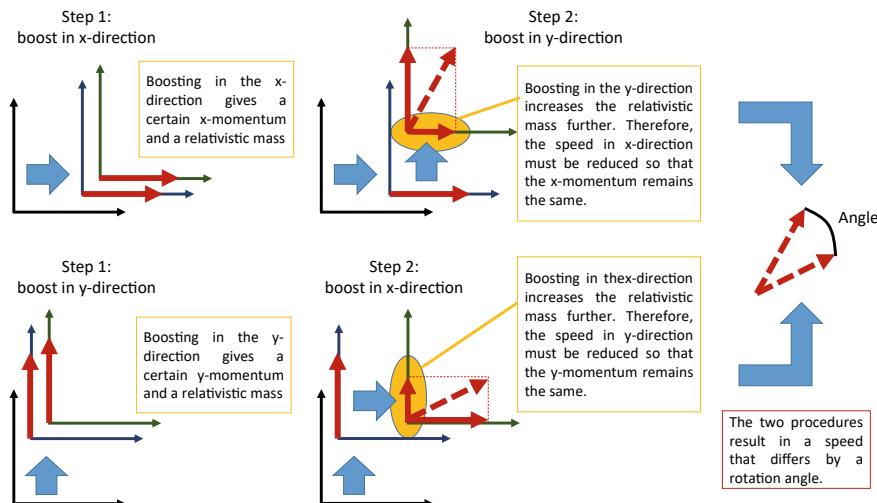


Fig. 13.5 A boost after another boost gives a boost and a rotation

This becomes interesting when we try to boost in the x-direction and then in the y-direction. If we do these operations one after the other, it turns out that the second boost is rotated. It becomes worse when we try to do it the other way around: first boost in the y-direction and then in the x-direction. In this case the second boost is rotated the other way. The consequence is that boosting x and then y gives a different result than boosting y and then x (see Fig. 13.5).

What could be the cause of this effect? Suppose we boost in the x-direction. The other system gets a velocity and momentum in the x-direction. When this velocity is close to the speed of light C, the relativistic mass of the object will increase.

Now suppose we boost it in the y-direction. The system gets a velocity and momentum in the y-direction. However, when we do that, *the relativistic mass of the system increases even further*. During this second boost, the momentum remains the same in the x-direction, but the relativistic mass increases. This can only be so if the velocity in the x-direction slows! So, we have the strange effect that boosting in the y-direction reduces the velocity in x-direction. When we do this the other way around, we get the opposite effect (see Fig. 13.5). The difference is a rotation about the z-axis. This rotation is called “Thomas rotation” after the physicist Llewellyn Thomas who lived in the twentieth century and described the impact of this effect on atoms and elementary particles [Ref. 14].

Another way of seeing this is by realizing that in relativity, a boost is actually a rotation between space and time! This can be seen in any Minkowski diagram (see Fig. 6.8): when you have zero velocity, you follow the time axis of the diagram. When something gets a velocity relative to you, it starts to follow a line that makes an angle with the time axis. So, a boost in one direction followed by a boost in another direction can be compared to a rotation about one axis followed by a rotation about

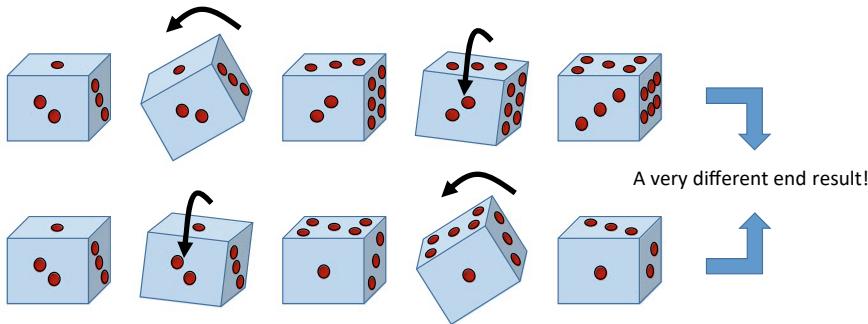


Fig. 13.6 Rotating about one axis followed by another gives a different result compared to the other way around

another axis. Rotations do not give generally the same result when they are carried out in a different order. Take for instance a die. When you rotate the die about one axis and then about another, you get a different result compared to when you do it the other way around (see Fig. 13.6).

In fact, when you take a round trip (boost in the x-direction, followed by a boost in the y-direction, then back in the x-direction and back in the y-direction), you find yourself at rest compared to the original situation, but you will have rotated!

The implication of this is that these linear transformations, the boosts, do not form a group in relativity. Remember from our discussion of symmetry groups that an operation belonging to such a group gives a result that is symmetric with respect to the starting point. For instance, the triangle looks the same after a rotation of 120° . We can also perform another group operation and get back to the starting point. From the previous discussion it follows that we can apply two boosts and end up with a rotation, an element that does not belong to the set of boosts and/or does not lead us back to the starting point. The consequence is that we have to consider a bigger set, that of boosts *and* rotations, in order to get a group. This is called the Lorentz group. If you also include space–time translations (i.e., change your position in space and/or time) you get an even larger group: the Poincaré group.

Summarizing, the Poincaré group consists of (combinations of) three operations:

- Boosts
- Rotations
- Translations.

It describes the fundamental symmetry of space–time in special relativity. It takes all three operations together to get the symmetry right. Consequently, *space–time requires us to consider that any object may be rotated relative to us*. It just takes a boost or two to make this happen.

When we consider waves, what does it mean to say that a wave may be rotating? Basically, a rotating wave is a helix (see Fig. 9.9). So, we can in fact make a wave move in a circular way. This can be described by a plain circle. However, as it turns

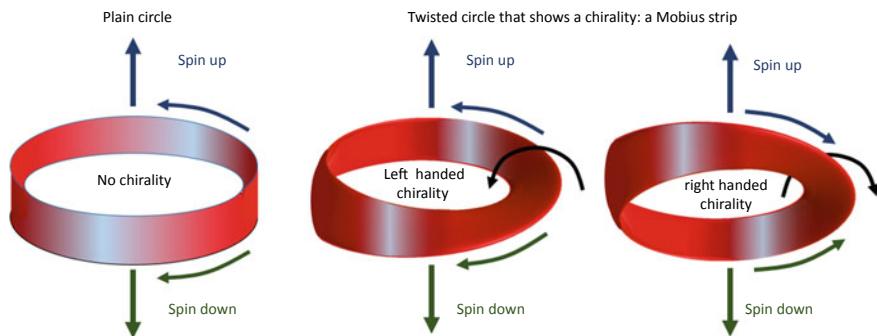


Fig. 13.7 Depicting a rotation as a simple circle (left). A wave can rotate along a circle in two directions (spin up and spin down). When we add another rotation on top of the circular motion (or rather, a twist), we get a Möbius strip. The twist can be one way around (left-handed chirality) or the other way around (right-handed chirality)

out this is not the whole story. On top of the rotation, we can imagine a second rotation. An object in a circular motion can also move *around its orbit*, rather as a planet turns around the sun as well as around its own axis. We can view this type of rotation as a rotation around a strip that is twisted. Such a twisted strip is called a Möbius strip. So, we have two types of rotation (see Fig. 13.7).

When a wave is helical as on the simple circle, it can rotate in two directions. These two directions can be viewed as spin up and spin down. However, when a wave rotates in the Möbius way, there are four options for rotation: on the middle Möbius strip in Fig. 13.7 with spin up and spin down as well as on the right-hand Möbius strip with spin up and spin down.

These two versions of the Möbius strip correspond to what is called the chirality of the wave function, as we will discover later. The first two spin states are referred to as having left-handed chirality, while the second two spin states are said to have right-handed chirality. So, we end up with the following four possible states:

- left-handed + spin $+1/2$
- left-handed + spin $-1/2$
- right-handed + spin $+1/2$
- right-handed + spin $-1/2$.

A wave that has this property of a twisted rotation around a Möbius strip is called a spinor. We will discuss spinors later.

We can conclude from all this that including special relativity in our quantum mechanical description of particles leads to waves having a spin state as well as a chirality. Both spin and chirality are therefore properties of a wave that originate from the structure of space-time as described by special relativity. They are a direct consequence of the Minkowski metric of space-time.

We can say that such a particle is Lorentz symmetric. This means that, when a Lorentz transformation is applied to it, its physical properties do not change. For

instance, when we describe the object from a different frame of reference (an observer going at a different velocity), it still looks the same. This is an important requirement. If a particle did not act conform to this requirement, it would violate the symmetry of Minkowski space–time. One consequence of that would be that we could potentially distinguish between different frames of reference and we could appoint a “preferred” frame of reference based on that distinction. That would imply that there is something like an absolute speed, and this would go against the spirit of relativity.

The propagator will have to be modified to include spinors. When that is done, the propagator describes whether a left-handed part or a right-handed part is entering or leaving the Feynman diagram. For each situation there is a different propagator that includes that particular state. A left-handed particle entering and leaving may have switched to right-handed in between, multiple times, or not at all. We will see later when we discuss chirality that a massive fermion can oscillate between left- and right-handed.

Both spin and chirality have some special properties. Spin is related to helicity, which is also a property of these particles. So, let’s dive deeper into spin, helicity, and chirality.

13.1.3 Spin as a Wave

We have learned that the relativistic properties of the vacuum require anything to be able to rotate or rather spin. The spin of a particle could be compared to the orbital momentum when we view the wave in a circular motion, e.g., such as a helix. In that case, the wave does not only wave along a straight line; it can also wave perpendicular to the line in a circular motion (see Fig. 13.8).

In the light of a particle being rather a cloud of disturbances than a single wave, it might be more correct to view spin, not as a wave becoming a helix, but as the virtual cloud getting a spin. Since charge was distributed in this cloud as well (vacuum polarization), we can imagine that such a spinning cloud can have a magnetic moment. For the sake of simplicity, we will be sticking to the model of a single wave. Some concepts relating to spin will otherwise become difficult to explain.

Now let’s view the spin as the part of the wave that goes around the circle (the circular part of the helix). Now we can fit that wave on the circle (see Fig. 13.9). This part of the wave that goes around the circle is not related to the particle’s velocity, momentum, or frequency. It is a fixed part that belongs strictly to the type of particle. For example, a photon has spin 1, which means that it fits one wavelength on the circle, just as in Fig. 13.9.

We assume that there is only one type of circle, with one radius. This would be dictated by the vacuum and valid for all elementary particle waves. Each field associated with a particle inherits this circle from the vacuum. It represents the extra degree of freedom that each wave can move in. We may also say that the virtual spinning cloud around the particle defines (the average radius of) this circle, but how that should be calculated I would not know! In any case, keep in mind that the math

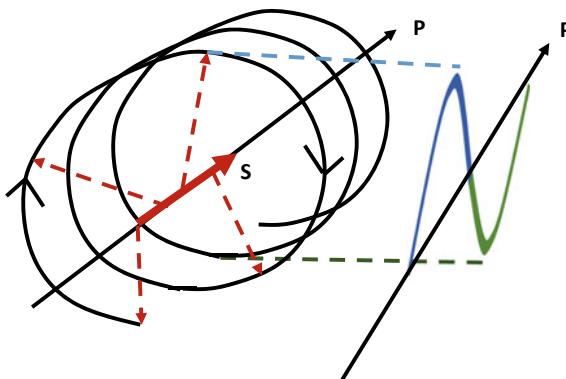


Fig. 13.8 When a wave has to spin, it turns into a helix. The helix contains a wave part along the axis of motion (the blue and green linear wave part on the right) and a circular part that makes a certain angle with the axis of motion (the red spin wave part). For example, the spinning can be perpendicular to the axis of motion. In that case (as in the picture), the spin vector (red arrow) that indicates the spin direction points in the direction of motion (black arrow) or opposite to the direction of motion (in which case it would spin the other way around)

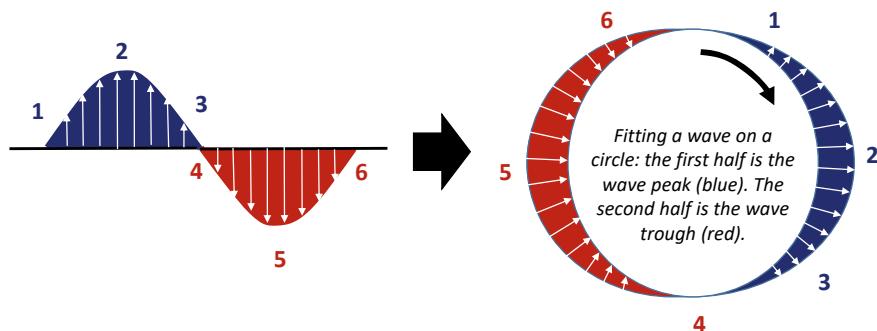


Fig. 13.9 Fitting a wave on a circle. The numbers on the wave agree with the numbers on the circle so that you can see how the wave fits on the circle

does not give any clue about how to view this, so the view presented here is just one of the possible views that might explain spin.

We can fit 0, 1, 2, ..., n waves on such a circle. Each value of n corresponds to spin n, so spin 1 fits exactly one wavelength on the circle. Spin 2 fits exactly two waves on the circle (see Fig. 13.10). For that wave, its circular wavelength is half as long and its circular momentum is therefore twice as big. One could view this as a dynamo determining the magnetic moment of the wave.

However, this is not all. We can fit another type of wave on this circle! We can fit a wave that takes two trips around the circle to complete one wavelength. This is a peculiar situation, since it is not allowed in the orbital momentum case. We will see later why this could be allowed in the case of spin.

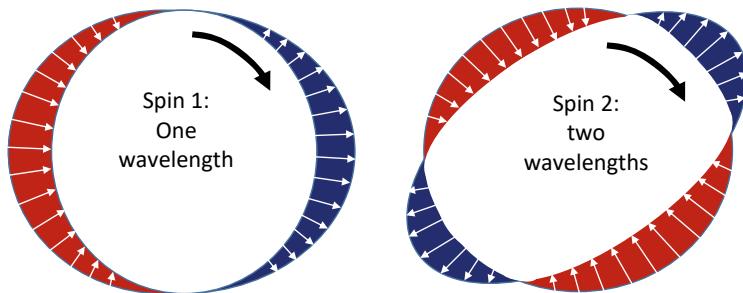


Fig. 13.10 Examples of spin 1 (one wavelength fitted in one circle) and of spin 2 (two wavelengths fitted in one circle)

Let's investigate this a little further. Such a wave will have completed $1/2$ of its cycle after going around (see Fig. 13.11). We can fit $1, 3, \dots, 2n + 1$ waves on a “two times around the circle” trajectory. One wavelength going around twice uses up half a wave to go around once (as before). Two wavelengths going around twice would be like the original “one wavelength on one circle”, and we have this type already (spin 1). The next *unique* way that requires two round trips to get back to its original state is three wavelengths going around twice (see Fig. 13.12). In that case, the wave has completed $1\frac{1}{2}$ waves on the first round trip. So, we end up with $1, 3, \dots, 2n + 1$ waves going around twice, or $(2n + 1)/2 = n + \frac{1}{2}$ per round trip.

We end up with two types of wave:

1. Waves that can spin n times on the circle.
2. Waves that can spin $n + \frac{1}{2}$ times. We will see later why this is allowed, in contrast to the orbital momentum example.

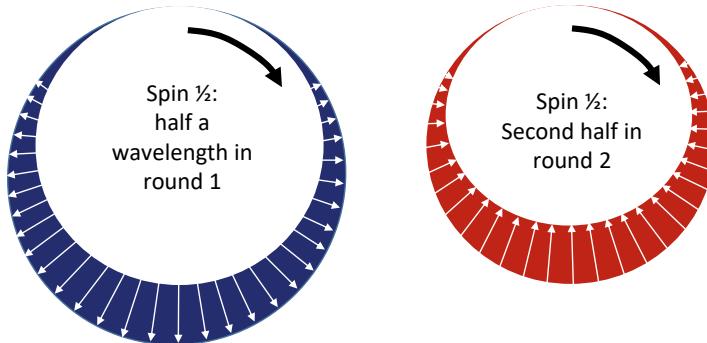


Fig. 13.11 Example of spin $\frac{1}{2}$. On the first round trip we see the first half of the wave: the wave peak (blue). On the second round trip we see the second half of the wave: the wave trough (red)

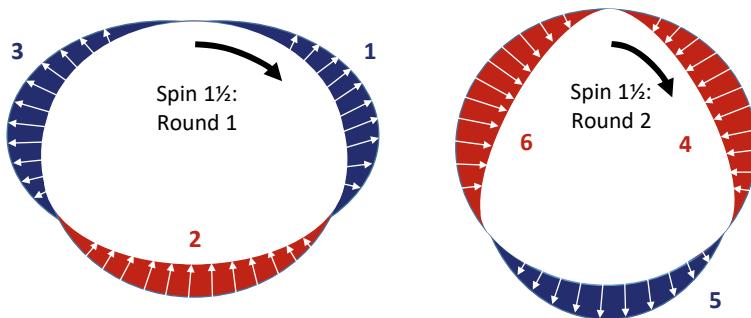


Fig. 13.12 Example of spin 1½. On the first round trip we see a wave peak (1), a wave trough (2), and a wave peak again (3). On the second round trip we see a wave trough (4), peak (5), and trough again (6) and we are back to the starting point. Clearly, this wave has completed 1½ wavelengths on one round trip

13.2 Fermions and Bosons

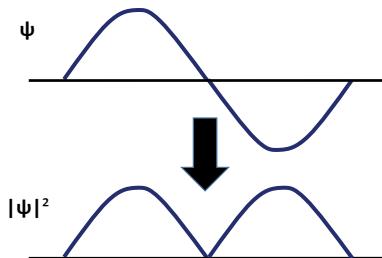
Let's look at the probability of finding a particle in a given state. The probability depends on the wave amplitude, or the field strength. As we saw before, a strong field implies a high probability of finding the particle there.

In quantum mechanics, the probability is found by taking the absolute square of the wave function (according to the Born rule). This means that the wave is squared (see Fig. 13.13). This makes sense, since a wave has negative values half the time and a probability cannot be negative. So, by squaring the wave function, we get a high probability where the field is strong (negative or positive).

When we describe a particle's wave function with momentum p as $|p\rangle$, the probability P is equal to $P = |\langle p\rangle|^2$.

Now suppose we have *two* states that fit a particle wave. We put one particle wave in one state and a second in the other. The resulting wave function looks like $|p_1 p_2\rangle$, where p_1 stands for the first particle and p_2 for the second. You can also put the second particle in the first state and the first particle in the second state. This looks like $|p_2 p_1\rangle$, with the particles exchanged.

Fig. 13.13 The Born rule: the wave squared gives the probability. The probability is always positive, and its magnitude is larger when the wave amplitude is larger



When the two particles are indistinguishable, the probability of the resulting “two particles in two states” must be the same, no matter whether you put p1 in state 1 and p2 in state 2 or vice versa. So, this means that

$$|\langle p_1 p_2 \rangle|^2 = |\langle p_2 p_1 \rangle|^2$$

For this equation to be true, the two wave functions must be equal or opposite:

$$|\langle p_1 p_2 \rangle| = +/ - |\langle p_2 p_1 \rangle|$$

These two options describe what we call symmetric and anti-symmetric wave functions:

$$\text{Symmetric: } |\langle p_1 p_2 \rangle| = |\langle p_2 p_1 \rangle|$$

$$\text{Anti-symmetric: } |\langle p_1 p_2 \rangle| = -|\langle p_2 p_1 \rangle|$$

13.2.1 Wave Phase

The above demands that two particles must get together in either a symmetric or an anti-symmetric wave function. There are many ways to combine two of the same waves into either of those. Let’s take two particle waves getting into a *symmetric* wave function together. When particle wave 1 is in phase 0 (the wave just starts to go up, like the start of a sine wave), wave 2 may be in any other phase. Now if wave 2 were in just the right phase to make it go exactly opposite wave 1, they would cancel each other out. This will not work. However, this situation looks like an *anti-symmetric* wave function. The fact that the two particles were in a symmetric wave function together rules out such a situation! Therefore, we can say that when two particles get into a superposition and form an (anti) symmetric wave function, *they must accord their phases before they do so*.

13.2.2 Fermions

Let’s look at the anti-symmetric case: the exchanged version is minus the unexchanged version. How can a combined wave be minus its exchanged version? This can only be when the combined wave itself is antisymmetric (see Fig. 13.14). This in turn means that the particle wave for p1 must be the same as the particle wave for p2, but opposite. And this can only be when both particles are indistinguishable, for otherwise the waves cannot be each other’s opposite.

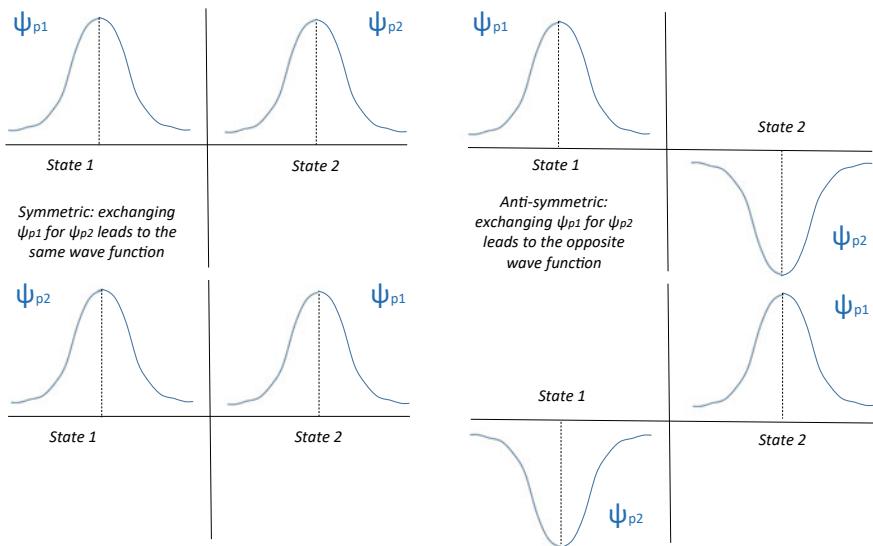


Fig. 13.14 When the two particle waves for p_1 and p_2 are exchanged, the symmetric wave function changes into itself (hence symmetric), while the anti-symmetric wave function changes into its opposite. This can only be when the waves for p_1 and p_2 are essentially the same, but each other's opposite, and this only works for indistinguishable particles. Both the symmetric and anti-symmetric wave functions have the same probability: the squared wave functions will all be positive and of the same shape

Now suppose, we have just one state to put a particle in. We still consider the anti-symmetric case. The wave function for p_1 is exactly the same as the wave function for p_2 , but they are each other's opposites. Only now, they are in the same state so they must interfere with each other. Because they are each other's opposites, they exactly cancel each other out. This means that the net probability of finding p_1 and p_2 in this state is 0. Hence, these two identical particles with opposite wave functions can only exist in different states, but never in the same state.

How can two identical particles have opposite wave functions? This must mean that they actually do differ in some way, while still maintaining their indistinguishability. Position and momentum are distinguishable if they differ. So is spin. However, two particles can have the same spin (and spin direction) while having opposite wave functions in the case of spin $\frac{1}{2}$. Why is that?

We saw before that waves of spin $\frac{1}{2}$ have to go round the circle twice before arriving at their original position. When such a wave is on the first trip round, it is actually exactly opposite to when it is going through the second (see Figs. 13.11 and 13.12)! We will go more deeply into this property when we discuss spinors. For now, we need to understand that particles with spin $\frac{1}{2}$, $\frac{5}{2}$, $n + \frac{1}{2}$ will always be in an antisymmetric wave function together. These particles are called fermions.

It is this characteristic that makes it impossible for fermions to be in the same state, ever. All particles that constitute matter (as opposed to forces) are fermions.

The reason for that is directly related to this spin characteristic. For instance, in an atom we have a limited number of states (as we saw before). These states can all be occupied with electrons in two spin states (spin up and spin down). Any more will interfere negatively leading to a 0 probability for the electrons to be in the same state. So, the only way for more electrons to fit into the atom is to occupy states further away from the atom's nucleus. One consequence of this is that matter takes up space, that atoms will not occupy each other's environments, and consequently that ever more matter will pile up into sizeable objects such as planets and stars. So, these characteristics of matter originate from the spin $\frac{1}{2}$ kind of behaviour of fermions.

13.2.3 Bosons

Bosons are the symmetric type of particles. They have spin 0, 1, 2, etc. So, these go around the circle once or twice or more. No matter how we try, we cannot get these wave functions to interfere negatively. Let's see how this works. We start from the general principle that two waves in a combined state start with the same phase. With one spin direction, the wave goes up (upturn) into the spin direction. If we take another particle wave that spins in the other direction, the wave goes down into the other direction (see Fig. 13.15). However, such a wave is exactly in phase with the first wave. So, bosons with opposite spin will always interfere positively with each other.

So, bosons are particle waves that always end up in symmetric wave functions together. This also means that, if we put two bosons in the same state, they will interfere positively with each other. Consequently, the field strength doubles. The probability for this is squared, so $2^2 = 4$. Hence, the probability for this to happen is greater than for these particles to remain in different states.

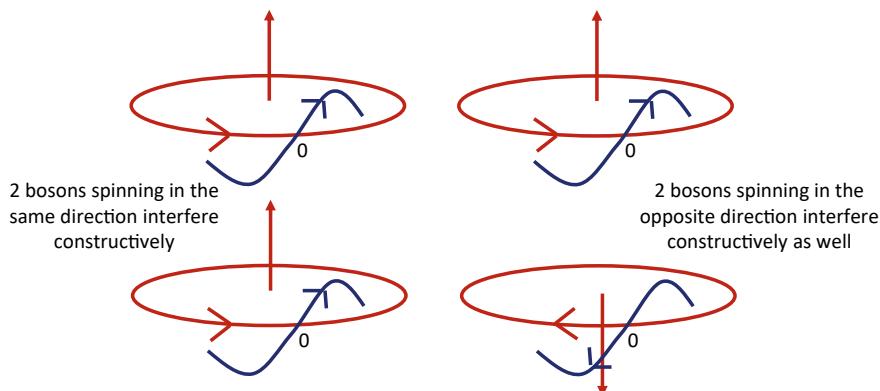


Fig. 13.15 Two bosons will interfere constructively, whether they spin in the same direction (and are in the same state) or whether they spin in opposite directions

This boson characteristic leads to the concept of the laser. In a laser, a certain type of atom is brought into an excited state. This is a state where electrons occupy a higher orbit than they could in the atom. These electrons will fall back into their original state and emit a photon of a typical frequency. Each atom that emits a photon will emit one at the same frequency, since this is determined by the energy difference between the orbits of the electrons. For the same type of atoms that is the same energy difference. The important thing is that the photons get close to other atoms that are still in an excited state. When they do, the probability for a photon to be in its “absorbed state” are twice as low as they are for it to be in its “emitted state”. Hence, the atom is stimulated to fall back and emit the photon. This is why it is called a laser: “Light Amplified by *Stimulated Emission of Radiation*”. So, the more photons there are around, the more atoms emit photons. This becomes a snowball effect, producing a very strong beam of photons. In order to amplify the effect, the photons are reflected between two mirrors so they pass the emitting atoms several times before they exit the laser. All photons leaving the laser are of the same frequency and they travel exactly in phase, so are concentrated in a narrow, highly focused beam. This gives us an extremely coherent beam of radiation.

Taking this example, it becomes clear why bosons can build up a coherent field: it is because they like to be in the same state. Moreover, we can put as many bosons in a small area as we want. Actually, they will like it, so there will be rather more than fewer bosons in that area. This means that bosons have no mechanism to build an object of any size as fermions do. Bosons are more of an extravert type of particle: the busier the better! Fermions are a rather introverted type of particle: they need their own space and prefer not to be swamped within a large group.

Consequently, it takes bosons to get a concentrated coherent field that has enough power to impact, e.g., fermions on a continuous basis. This is what is needed to be able to bend the paths of fermions in any significant way. So, bosons typically carry forces, while fermions create sizeable matter. The interplay of fermions and bosons creates larger structures: fermions for the size of it and bosons to make them stick together. A bit like bricks and mortar.

13.2.4 *Spinor Fields*

We saw that we cannot shift phase on the circle, resulting in bosons always interfering positively when they are in the same state. Fermions interfere negatively because they are in the same phase on the circle, but on the other trip round. You might argue that this can be seen as a phase shift of one round trip. Why would that be allowed? Moreover, fermions differ in this characteristic compared to orbital momentum. Why can we not do this trick for orbital momentum?

In order to answer this question, we have to look at the Möbius strip (Fig. 13.16). We create a Möbius strip by taking a regular straight strip, turning one end through 180° around an axis along the strip, and connecting it to the other end. Then, when we go around once, we find yourself on the other side of the strip. We would have

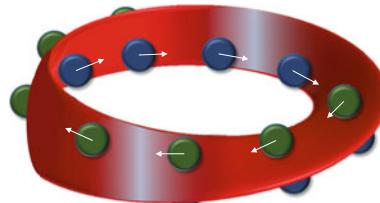


Fig. 13.16 Möbius strip. If we follow the green ball on the first trip around the strip, we find ourselves on the other side of the strip. The blue ball indicates the second trip around the strip. The blue ball ends where the green ball started, showing that it takes two trips round to get back to the original position

to do one more trip round to arrive back at the original side again. We can view spin $\frac{1}{2}$ particle waves as waves that use one wavelength around the Möbius strip once, implying that they would have to go around twice to get back at the original spot. This way, there is no phase shift. It is rather a rotated path. Looking at it this way we do not have to worry about phase shifts suddenly being allowed for fermions.

From this picture we can conclude that fermions are waves that spin on Möbius strips, while bosons are waves that spin on regular strips. So now we can connect Fig. 13.7 and the two types of rotation that are depicted there with bosons and fermions.

Using this picture, we can explain why there are only two types of spin wave: on Möbius strips and regular. Suppose for instance that we cut the strip and twist it twice instead of once. You would get a “twice twisted” Möbius strip. But when we go around this, we stay on the same side, so we can only go around the circle once. When we turn it three times, we end up on the other side and we have the same experience as with the original Möbius strip. Therefore, there are only two types of situation: regular and Möbius.

That’s a great idea, the Möbius strip, but how should I view a Möbius strip in the vacuum? Of course, this is not an actual strip (as it would have to be “made of something”). So take the picture of the wave going around the Möbius strip twice and take out the Möbius strip itself. What remains is simply a wave that makes half a twist on going around a circle. It completes that twist on going around the second time. That’s all it is. And now we can connect this to the earlier conclusion that relativity requires our being able to rotate in space in two different ways: the twisted way and the non-twisted way.

So, spin differs from orbital momentum in that the wave may twist while going around the circle. Spin $n + \frac{1}{2}$ particles (fermions) make such a turn. Spin n particles (bosons) do not.

Let’s dive a little deeper into this picture. In Fig. 13.17 a wave is shown on a Möbius strip. In this picture we show one wavelength doing one trip round, as a boson would do. We start with a wave peak (blue spikes, connected by a blue line) followed by a wave trough (green spikes). The blue part is on one side of the strip since the wave peak is usually above the line. The green part appears on the other side

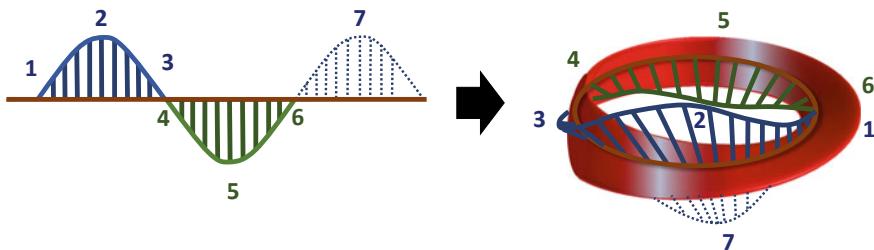


Fig. 13.17 A wave with one wavelength on one round trip. Numbers 1, 2, and 3 (blue) indicate the first peak of the wave, above the line and so above the strip. Numbers 4, 5, and 6 (green) indicate the trough of the wave, below the line and so below the strip (other side of the strip). When the second peak begins, it must switch sides again compared to the previous trough. Therefore, it must appear on the other side of the strip compared to the first wave peak! This is a consequence of the fact that the strip itself turns through 180° . Such a wave would clearly cancel itself out on the second round trip

of the strip since a wave trough is usually below the line. It becomes interesting when we look at the second round trip taken by the wave. The second round trip would start again with a wave peak, but now on the other side of the strip (blue spikes on the other side of the strip, also connected by a blue line)! This wave interferes negatively with itself! So, this wave would cancel itself out.

Now let's try a wave on the Möbius strip that takes two round trips to complete one wavelength. That means that the wave peak would take the entire first round trip and the wave trough the entire second round trip. This looks like Fig. 13.18.

One conclusion from this picture is that since the peak of the wave and the trough of the wave are in line, they amplify each other. They interfere positively. So, this wave can exist on the Möbius strip.

So, waves that go around twice in one wavelength cannot exist on a regular circle, for they would cancel themselves out. But they can exist on a Möbius strip. Clearly, these waves belong to fermions. A wave that goes around once in one wavelength

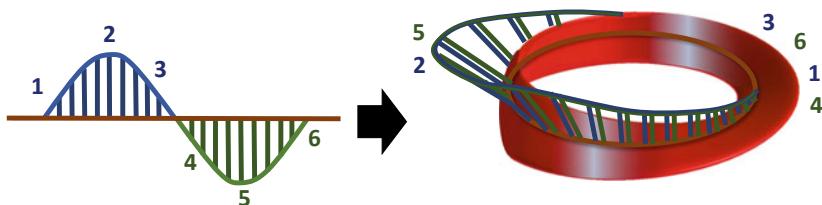


Fig. 13.18 When a wave goes around twice, another surprising thing happens. Numbers 1, 2, and 3 and the blue spikes indicate the peak of the wave. It takes an entire round trip to complete the peak of the wave. The peak ends at the other side of the strip. Consequently, the trough (green) must switch sides (the wave on the left moves from above the line to below the line). The result is that the trough goes exactly in line with the peak of the wave during the previous round! The trough is indicated by the green spikes and the numbers 4, 5, and 6

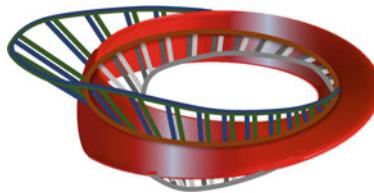


Fig. 13.19 Opposite wave functions on the Möbius strip. The green and blue spikes indicate the field strength (amplitude) as in Fig. 13.18. The dark and light grey spikes represent the amplitude of the opposite wave. Since there is no reason why the peak of the wave should start on one side or the other, the grey wave is just as likely as the green and blue one

cannot exist on the Möbius strip, for it would cancel itself out. But it can exist on a regular circle. These waves represent bosons. Summarizing:

	Regular circle	Möbius strip
Bosons: wavelength fits on one round trip	Amplifies itself	Cancels itself out
Fermions: wavelength fits on two round trips	Cancels itself out	Amplifies itself

Now it becomes clear why there can be only two types. Suppose you can twist the band by an arbitrary value instead of the 180° of the Möbius strip. Then the wave would not exactly reinforce itself and after a number of turns, it would start to become an opposite wave and cancel itself out. So, waves can only end up in a reinforcing run around a circle and around a Möbius strip. All possibilities in between will cancel themselves out eventually.

A surprising fact about fermion waves is that these waves have a field amplitude that stays only on one side of the Möbius strip! There is no reason why the wave should be on one side of the strip or on the other. In either case, however, the wave would stay on the chosen side of the strip.

This, in turn, has an interesting implication. It means that such a wave could exist in two types. Let's put both types next to each other in Fig. 13.19. Here we see that these two types of wave are opposites! So, we now have a basis for our earlier conclusion that two fermion waves are opposite. Their being opposite implies that a superposition of two fermions can be an anti-symmetric wave function.

Keep in mind that this analysis does not tell us why two fermions *must* be opposite. It just shows that they *can* be, while bosons cannot (unless you allow phase shifts on the spin circle). It is not clear why two fermions in superposition would get into a wave function where one would be on one side of the strip and the other on the other side. It should also be kept in mind that this view of wave functions on the Möbius strip is just a way to understand better what happens. In reality, the way it works could be different. However, mathematically it will be consistent with the view presented here.

We can conclude that the vacuum gives all fields either the same circle or the same Möbius strip to spin on. Different fields can spin with different values. On the

Mobius strip these values will be of the form $n + \frac{1}{2}$. On a regular circle they will take on an integer value n .

The fields that are defined on the Mobius type of circle are called spinor fields. The particle waves associated with this are called spinors. Spinors are waves that get to their original position by turning around twice (i.e., 720°). Consequently, all fermions are spinors. The Mobius strip is one way to see this, but there are other ways. Dirac's belt trick is a favourite amongst physicists [Ref. 8, p. 141]. You can also visit Wikipedia (spin or spinor), where you will find a beautiful film that shows how a connected ball can keep turning without twisting the ties it is connected to. This is just another way of showing how it can take two round trips to get to your original position. Another famous example is the Balinese belly dancer. This dancer keeps turning her hands around while twisting her arm one way (first round) and then the other way (second round).

You can try this out yourself by placing a cup on your hand and trying to turn your hand around in the same direction. When you start with your right hand up, you can turn the cup to the left by letting your hand go down, then turn under your arm. After that you turn your hand further while moving it up. So far, you have made one full turn while your arm feels twisted maximally. Now you go on by moving your hand up and over your arm, effectively twisting your arm back, then further around until you are back in the starting position. So, you see that you first did one full turn under your arm until your arm was maximally twisted and then you did a second turn over your arm, twisting your arm back again. You can keep turning the cup around, while your arm twists one way and back all the time. The first round is like one side of the Mobius strip. The second round is like the other side of the Mobius strip.

So, we have different types of fields related to spin:

- For bosons: scalar fields (spin 0, e.g., the Higgs boson) and vector fields (spin 1, e.g., gauge fields such as the photon)
- For fermions: spinor fields.

These fields borrow an internal degree of freedom from the vacuum that we can view as a circle or Mobius strip to fit a spin wave on.

13.2.5 Fermions in Opposite Spin

We know that in any state, we can fit one fermion with spin up and one fermion with spin down. How can two fermions of opposite spin be in one state together? The idea is that they are not in the same state when their spin differs. So, we may fit a spin up and a spin down fermion into one state. But how does this work when we look at it from the wave perspective?

Fermions in each other's neighbourhoods will get into an anti-symmetric wave function, so each of them will be spinning on the opposite side of the Mobius strip. If they had the same spin, they would cancel each other out.

However, what is the situation when their spin is opposite? A spin wave starts with the upper part of the wave in the spin direction. So, when spins are opposite, their spin waves will start on opposite sides of the Möbius strip. Since the fermions must be in an anti-symmetric wave function together, one of them starts on the other side of the Möbius strip. The result is that both fermions are now spinning on the same side of the Möbius strip. Consequently, they interfere positively and can occupy the same state together. Any more fermions would not fit since they would start to interfere negatively with one of the others.

13.3 Helicity

Helicity is nothing more than the projection of the direction of spin onto the direction of motion. The direction of spin is described by an arrow perpendicular to the surface generated by the spin circle. The direction of the spin arrow is determined by the right-hand rule. This rule states that, when you bend the fingers of your right hand into the direction an object is spinning, your thumb will point in the direction the arrow should be drawn. Thus, when you see a spin arrow, the only thing you have to do is point the thumb of your right hand in the direction of that spin arrow and you know the direction the object is spinning around. Obviously, there are only two possible ways to spin around a given axis: clockwise when looking along the axis (formally called spin down) or anticlockwise (formally called spin up) (see Fig. 13.20).

All this is just a matter of convention. There is no hidden meaning behind the right-hand rule, just a standard way agreed by physicists to make sure we all know what we mean by “spin up” and “spin down”.

With two spin directions, helicity also comes with two possible values:

1. Positive helicity = spin up in direction of propagation (hence spin down in opposite direction)
2. Negative helicity = spin up in opposite direction (hence spin down in direction of propagation).

We will see later how helicity relates to chirality.

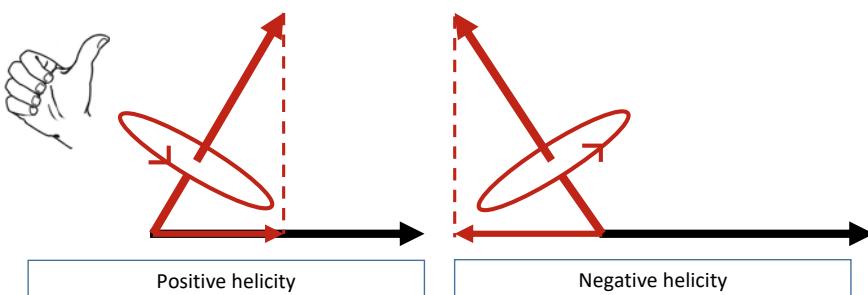


Fig. 13.20 The direction of spin. The arrow indicates the direction of spin. The arrow can be projected on the direction of motion. This will tell us the helicity of the spinning object

13.4 Chirality

We introduced chirality as the twist direction of the Möbius strip (see Fig. 13.7). However, we have not yet discussed its formal definition.

What is chirality? Chirality is the property an object has when it has a mirror image asymmetric to itself. For instance, a left shoe. In the mirror a left shoe looks like a right shoe. Suppose you could grab through the mirror, get the mirror image of the left shoe, and pull it out of the mirror (of course this is impossible, but suppose you could). Then you would be holding a right shoe. The right shoe cannot be turned in any way in three dimensions to become a left shoe. So, the shoes are not symmetric with respect to their mirror image. Shoes come in two chiralities: left and right.

Chirality originates from the Greek word “cheir”, which means “hand”. Your hands are a perfect example of an object that has two chiralities: left and right. You cannot rotate your right hand in any way so that it becomes your left hand.

So, a left shoe cannot be turned into a right shoe, or can it? Try as you will, in three dimensions you will not succeed. However, one dimension more and you actually can rotate a left shoe and get yourself a right shoe! To understand this, we take refuge in a lower dimension. Just draw a right shoe on a piece of paper (Fig. 13.21) and then take it out of the flat space, and turn it in our three-dimensional world. You end up with a left shoe!

So how can a particle have a chirality? Particles are waves, not shoes. It does not seem logical for particles to have a certain shape that gives them a chirality like shoes. So how does that work?

13.4.1 *Fermions Come with Two Chiralities, Called Left and Right. Bosons Do Not*

The core difference between bosons and fermions is in their spin. As we have seen this implies that fermions have a spin wave going around a Möbius strip, while bosons have their spin wave going around a regular circle.

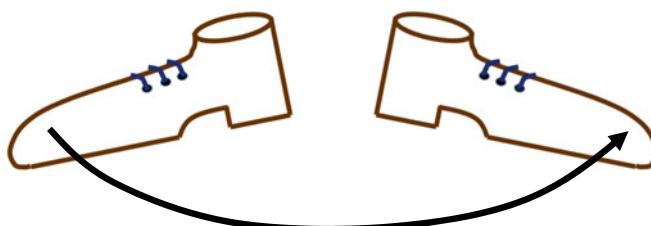


Fig. 13.21 A two-dimensional right shoe can be turned in our three-dimensional world into a two-dimensional left shoe. But if you try to do this by rotating the right shoe on the two-dimensional paper, you will not be able to make a two-dimensional left shoe out of it

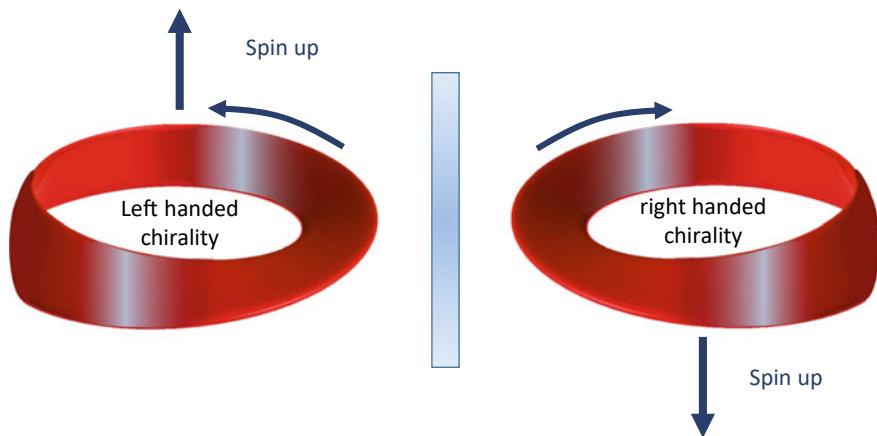


Fig. 13.22 The two chiralities of the Möbius strip

The question was: how can a wave have a chirality, like a left or right shoe? Since the difference between bosons and fermions is primarily in the Möbius strip, let's take a deeper look at it. The point is that the mirror image of a Möbius strip is again a Möbius strip, but one that cannot be rotated in three dimensions to be equal to its original (see Fig. 13.22). So, like shoes, a Möbius strip comes with two possible chiralities.

So how do these two chiralities of the Möbius strip differ? Let's look again at how to make a Möbius strip. You take a regular strip, cut it somewhere, turn one end through 180° and tape it back to the other end. Wait a minute! Turn one end through 180° ? You can turn it in two directions: left or right. When you turn one end through 180° left, you get one chirality of the Möbius strip. If you turn it 180° right, *you get its mirror image*, the other chirality! This is why you cannot simply rotate the strip with one chirality in three-dimensional space to get the other chirality. You must cut the strip, then turn one end 360° the other way and tape it back in order to get the other chirality. This cannot be done by making a rotation in three dimensions.

So now we know why we called the two twist-directions of a Möbius strip its two chiralities. They are each other's mirror images.

13.4.2 Under Parity, the Chirality of a Fermion Is Swapped to the Opposite Chirality

The parity operation is performed by taking the mirror image of an object and then rotate it through 180° around the axis perpendicular to the mirror. This operation represents a symmetry in space. One would not expect the laws of nature to change under this operation since a preferred direction in space does not exist.

If we do that with the Möbius strip, we see that we get the other chirality by taking the mirror image. The rotation through 180° moves it into an interesting position: it spins in exact the same position as before, while the Möbius strip is now “inside out”, corresponding to its mirror image. The direction of propagation is now in the opposite direction, which is a consequence of the parity operation (see Fig. 13.23). In the picture, keep in mind that it is the movements that are mirrored, not the arrows. For something that makes a trip around a circle, this means that the trip around the circle gets mirrored. When we describe that round trip with an arrow, the arrow points in another direction than its mirror image would. Therefore, the mirror reflection is responsible for the change in angle that the spin makes with the direction of motion, the original angle being α and the mirrored one $180^\circ - \alpha$.

So, we conclude that the mirror image is responsible for two effects that appear to differentiate left-handed chirality from right-handed chirality:

1. The angle with the direction of motion is different.
2. The Möbius strip twists in a different direction for the two chiralities.

As we will see later, the angle with the direction of motion is not the key distinguishing factor, since it may depend on the frame of reference.

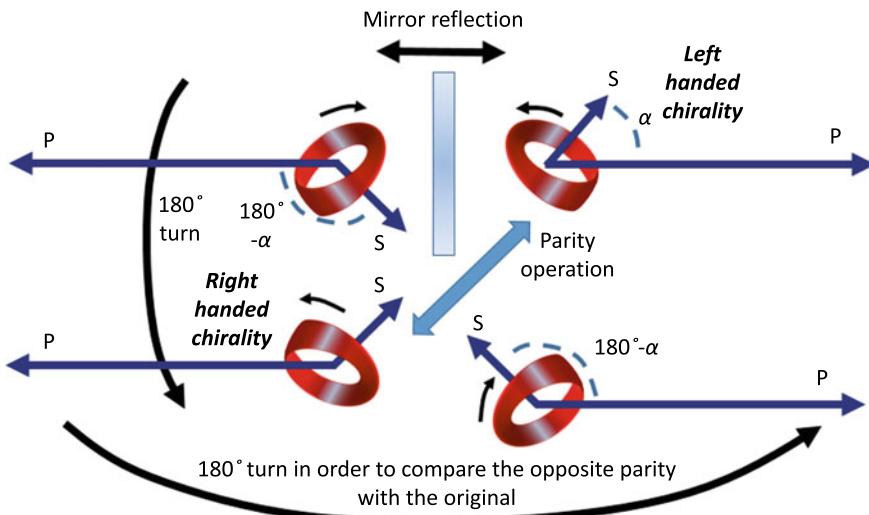


Fig. 13.23 Parity swap for a Möbius spin of right-handed chirality. After parity (mirror reflection + 180° turn), the spin makes another angle with the direction of propagation. In order to compare what happened with the original situation, the resulting left-handed chirality is rotated through another 180° . This leads to the version in the bottom right corner. This clearly shows that the left-handed chirality makes an angle of $180^\circ - \alpha$ relative to the direction of motion, while the right-handed chirality makes an angle of α with the direction of motion

13.4.3 Low Velocity Fermions Flip Chirality at the Frequency of Their Mass

In quantum field theory one can describe mathematically how fermions move. From this description one can show that at low velocities fermions swap from left-handed chirality to right-handed chirality with a frequency equal to their mass [Ref. 8, p. 326]. Others interpret this as “mass mixes the left- and right-handed states” [Ref. 30, p. 186]. But this is a strange situation. What should mass have to do with swapping (or mixing) chirality?

Let's recall what mass is. Mass is created by an extra potential. It leads to an extra frequency, the mass frequency. The Higgs field is the potential responsible for this. We compared the Higgs field to an endless set of springs pulling the wave. You could also compare it to gravity. What does gravity do to a spinning top? It creates a precession. When you spin a top and drop it on the floor, the axis is usually not straight up, but tilted. Gravity works on the tilted axis, pulling it down. The result of that action is that the axis moves sideways. In the end the axis turns in a circle around the perpendicular to the floor. This is called precession.

It is tempting to view the extra mass potential as performing a similar action on the spinning wave. The effect would be that the axis starts to precess. If the turning effect occurred in another dimension (not our three-dimensional world), the Möbius strip would swap from a left-handed to a right-handed chirality at a frequency determined by the mass. Just like the two-dimensional left shoe which we could turn into a right shoe by rotating in our three-dimensional world. This is pictured in Fig. 13.24. More generally, the interaction with the Higgs field swaps the chirality of a fermion. The more often it interacts with Higgs, the higher the mass (the stronger the springs), but also the more often it swaps chirality.

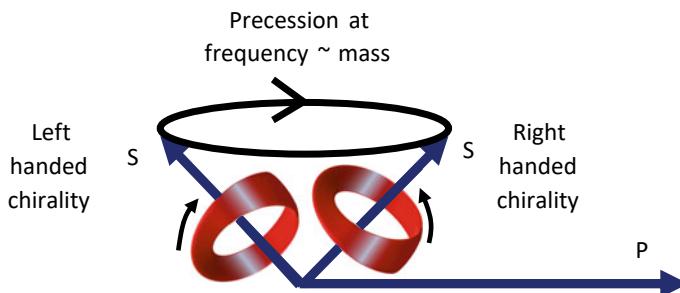


Fig. 13.24 Precession of chirality at the frequency of mass. Keep in mind that the Möbius strip is flipped in another dimension than our three-dimensional world, a dimension which we cannot picture here. So, this flip is different from the mirroring we showed in Fig. 13.23, just as flipping two-dimensional shoes in the third dimension is different from mirroring them in the flat space they were drawn in

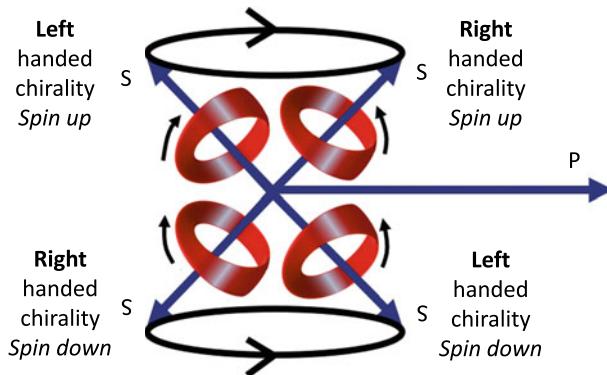


Fig. 13.25 Spin up and spin down in each chirality. Here we find all four solutions we also identified when we discussed the origin of spin

13.4.4 Chirality Is Not the Same as Spin

The two chiralities have the same effect: going around once leads you to the other side of the Möbius strip. So, the chirality of the Möbius strip has no impact on the spin waves. We can have spin up and spin down in both chiralities. Spin up is basically the spin wave turning in one direction (right way around), while spin down is turning in the other direction (left way around) on the Möbius strip. This can be done in each chirality in an equal manner. We can see this from Fig. 13.25, which indicates how spin up and spin down behave on each chirality.

The chirality–spin combinations correspond exactly to the four ways a fermion can rotate, as described in the section on the origin of spin.

13.4.5 Fermions of Different Chirality Are Different Particles

The twist in the Möbius strip is opposite for left-handed fermions compared to right-handed fermions. Consequently, the field amplitudes (the amplitudes of the waves) of left-handed and right-handed fermions point in different directions (see Fig. 13.26).

A spin wave in one chirality follows a different path with different directions of the spinor field compared to the other chirality. So, they cannot be each other's opposites and they cannot be in an anti-symmetric wave function together. The consequence is that we have to distinguish the two chiralities as two different particle types: left-handed and right-handed.

Does this mean that we can find a left-handed spin up electron and a right-handed spin up electron in one state together? No. Since fermions swap their chirality at the frequency of their mass, they cannot be in the same state. Maybe they would even align their “precessions”. Who knows? But in the quantum world we would not know

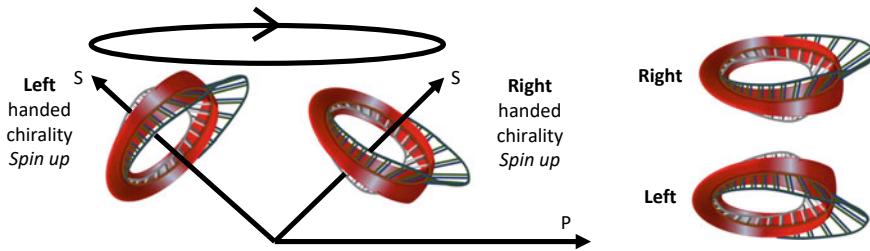


Fig. 13.26 Opposite wave functions in each chirality. All four possible wave functions point in a different direction, no matter what angle is made relative to the propagation. So, a spin wave in one chirality cannot cancel out a spin wave in the other chirality. Therefore, a right-handed fermion can only get into an anti-symmetric wave function with another right-handed fermion. Consequently, we must treat the two chiralities as different particle types: left-handed fermions and right-handed fermions

exactly which chirality a particle is in at any specific time. So, it can be in any of those chirality states and it will therefore interfere negatively with any other fermion of the same type.

The plot thickens when we realize that, in the standard model, fermions with different chiralities have different hypercharges and different isospins, as we will see later. Moreover, the weak force only couples to left-handed fermions. The result is that only left-handed fermions could change into their brother or sister in the standard model. We will get back to all this later.

So why would that be? This is an as yet unanswered question, but it does lead us to believe that the right-handed chirality way of spinning has a property (or specific reason) that makes it impossible to connect to the weak force. Massive fermions that swap from left to right-handed and back therefore only feel the weak force for part of the time, when they are in the left-handed chirality.

This also implies that nature is not symmetric under the parity operation. Just taking the mirror image of a left chirality particle changes the way nature treats the particle.

13.4.6 At Very High Velocities, the Chirality of Fermions Becomes Fixed and Related to Their Helicity

Massive fermions cannot move at the speed of light. Consequently, they can be overtaken by someone going faster. If the helicity of the particle is positive in a frame going more slowly, it will be negative in a frame going faster. This is a consequence of the definition of helicity: the projection of the spin direction onto the direction of motion. The direction of motion of the fermion is opposite in the two frames. However, the direction of spin remains the same. Just imagine a spinning top. When you pass by the top in the other direction, the “up” direction does not change. Suppose

that the direction the top spins around is linked to the “up” direction by the right-hand rule (see Fig. 13.20). When you pass it by from the other direction, you would still apply the right-hand rule to the “up” direction to describe the direction of spin. If the top were really spinning in the other direction, you would conclude that it was spinning in the “down” direction using the right-hand rule. This direction of spin is independent of the direction in which you pass it by.

One consequence of this is that, in a different frame, the direction of motion may well be different but the direction of spin will not be. Hence, the helicity will depend on the frame of reference (the relative velocity of the observer).

The chirality does *not* change when observed from different frames, since this would require a parity operation which includes a mirror reflection. A mirror reflection cannot be achieved by changing a frame of reference, which is just a different speed and direction of motion. So far, it is clear that helicity and chirality are very different things (see Fig. 13.27).

When the velocity is 0, we draw spin up as an arrow up and spin down as an arrow down. When the velocity goes up, the arrow starts to become more aligned with the direction of motion. This is a relativistic effect. When the spin makes an angle with the direction of motion, we see the spin motion as partly parallel to the direction of motion and partly perpendicular to it. The parallel part will suffer from length contraction at high relative velocity. Very close to light speed, the length contraction will have diminished the parallel part of the spin motion, so only the perpendicular part remains. The result is that the arrow of spin direction has turned so that it has become parallel to the direction of motion (see Fig. 13.28).

The angle of the right-handed spin up on the left in Fig. 13.28 would become equal to positive helicity at velocities close to c . On the right in Fig. 13.28 the angle of the right-handed spin up would become equal to a negative helicity.

However, the math shows that at very high velocities, a right-handed chirality particle has a positive helicity (spin in the direction of propagation) [Ref. 8, p. 329], while left-handed chirality has a negative helicity at very high velocity.

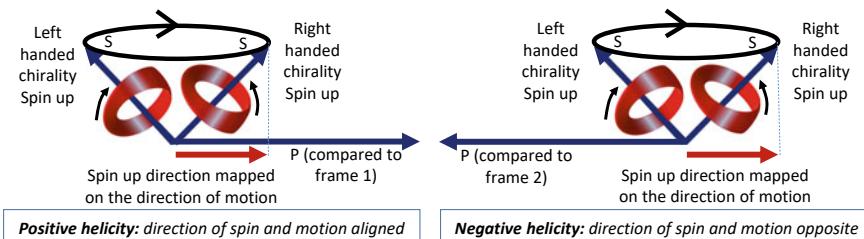


Fig. 13.27 Chirality and helicity. Neither spin direction nor chirality change when observed from a different frame of reference, but the direction of motion does. Hence, helicity depends on the frame of reference, but chirality cannot. The angle of the chirality with the direction of motion depends on the frame of reference, which leaves us with only one distinguishing factor between left- and right-handed chirality that is frame independent: the way the Möbius strip is twisted

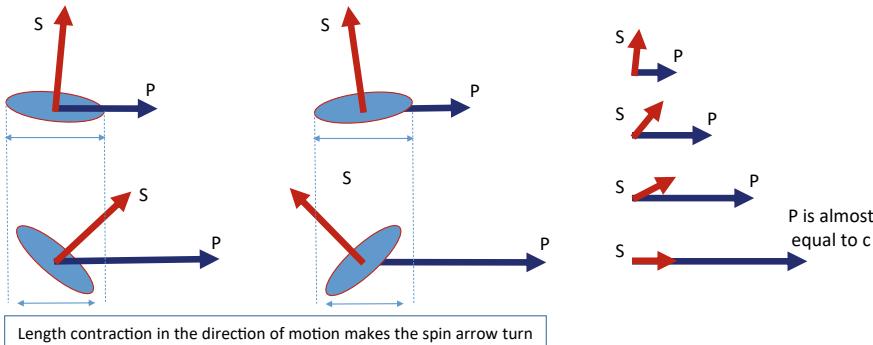


Fig. 13.28 When the velocity p increases, length contraction along the velocity direction makes the spin arrow turn. When the velocity is very close to c (the speed of light), the spin arrow will be close to parallel to the direction of motion

So apparently a particle with spin up in the direction of motion (positive helicity) selects the right-handed chirality as the preferred chirality. One step further: if this fermion were massless, it would have a fixed positive helicity since it cannot be overtaken. There is no frame which goes faster, so there is no frame in which the helicity can be swapped over. Such a fermion would also have only one fixed chirality: right-handed.

The same is true for a particle with spin down in a direction opposite to the direction of motion. This would select the left-handed chirality to become the preferred chirality. If this fermion were massless, it would have a fixed negative helicity and a fixed left-handed chirality.

How could the spin direction and the direction of motion select a chirality? This seems a bit fishy. The chirality is determined by the twist in the Möbius strip. This induces the spin wave to turn around the strip in a certain way.

At (almost) light speed we could argue that it is hard (or even not possible at light speed) for the field strength to turn in the direction of motion, as it would at that point go faster than the particle. At the speed of light this is not possible. So, let's take a look at the way the field twists around the Möbius strip (see Fig. 13.29).

The requirement that the field strength must turn away from the direction of motion at (almost) light speed does lead to the desired pattern as described by the math, i.e., it implies that spin up in combination with right-handed chirality makes the field on the Möbius strip turn away from the direction of motion. The same works with the opposite situation: left-handed chirality combined with spin down. The other combinations make the field turn in the direction of motion, which is not allowed. Consequently, right-handed chirality cannot exist in a negative helicity situation and left-handed chirality cannot exist in a positive helicity situation.

What does it mean that waves are not allowed to go faster? Clearly, we know this from relativity. However, using the wave picture we can make that a bit more specific. Waves in the field can propagate maximally at light speed. So, when the field tries to propagate faster, which is what would happen in the forbidden combinations, it

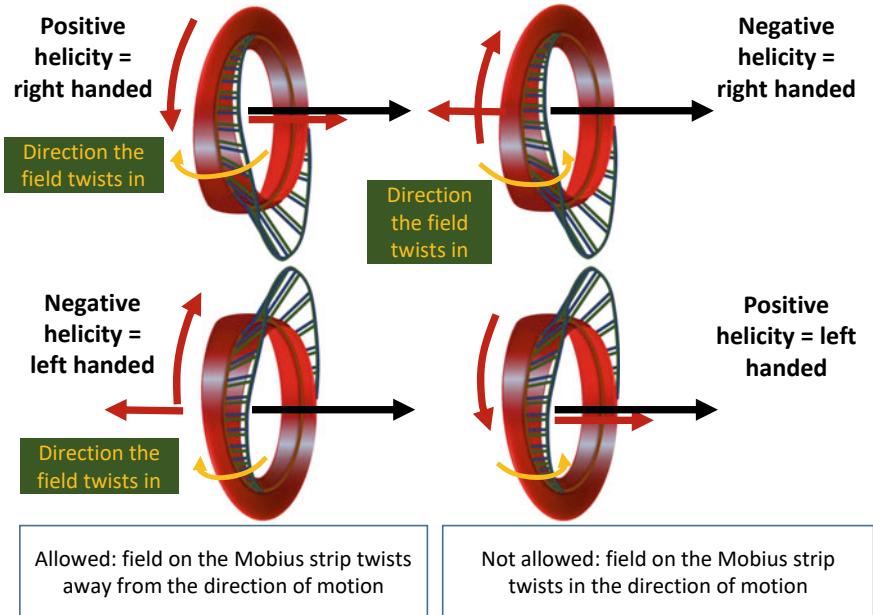


Fig. 13.29 The field strength turns away from the direction of motion in the combinations “right-handed + positive helicity” and “left-handed + negative helicity”. The other two combinations lead to the field strength turning in the direction of motion, potentially going faster than the particle, which is not allowed. So, e.g., “right-handed chirality + negative helicity” is not allowed

simply does not succeed in doing that. This implies that the field cannot sustain a spin wave in these situations. Consequently, such a state would cease to exist. All this results in the “allowed” combinations being the only ones we will see from a high velocity reference frame. Keep in mind that an observer in another frame of reference will still be able to see both chiralities. It is only for massless fermions that this will never be the case.

13.5 Fermions Becoming Bosons

Fermions can be combined into other particles. These composite particles can exhibit boson behaviour. Examples are mesons (consisting of a quark and an anti-quark) and He^4 (consisting of two protons, two neutrons, and two electrons). Mesons behave like bosons and, among other things, they carry the strong force between protons and neutrons in atomic nuclei. He^4 exhibits boson behaviour when it is cooled below 2.17 K. Indeed, it exhibits superfluidity, because the He^4 atoms get into the same state [Ref. 8, p. 370].

This implies that such a combination of fermions can get into a symmetric wave function. So how does that work? First of all, the total spin of the composite particle is the sum of the spins of its constituents. As a result, a composite particle made of an even number of fermions has an integer total spin. This in turn means that the composite particle has a symmetric wave function. For example, two anti-symmetric wave functions can get together and form a symmetric wave function, just as a left hand and a right hand (which are anti-symmetric to each other) are *as a pair* symmetric to another pair of left and right hands. Hence, such a *composite* particle acts as a boson. This means that, *as a composite particle*, it likes to be in the same state as its fellows.

The restriction is that this only works on the level of the composite particle. Two He^4 atoms cannot pile up in the same position for instance (like photons can). When one He^4 gets really close to the other He^4 , the individual wave functions of the constituents come into play and these are fermion wave functions. Consequently, He^4 has the same volume when it becomes superfluid. What these atoms can do is show coherence, and that is what we see when the fluid becomes superfluid.

Another aspect is that a composite particle must be de-localized to make its boson behaviour apparent (see Fig. 13.30). This means that a single He^4 cannot be localized in a certain position, so the symmetric wave functions of these atoms stretch beyond their “positions”. Only then can they interfere with each other’s total, symmetric, wave function and show boson behaviour. For particles such as He^4 , this means that their momentum must be very low. Only then is their position largely unknown according to the uncertainty principle. So, He^4 exhibits superfluidity only at very low temperatures, i.e., temperatures low enough to allow the wave functions to be sufficiently de-localised to overlap.

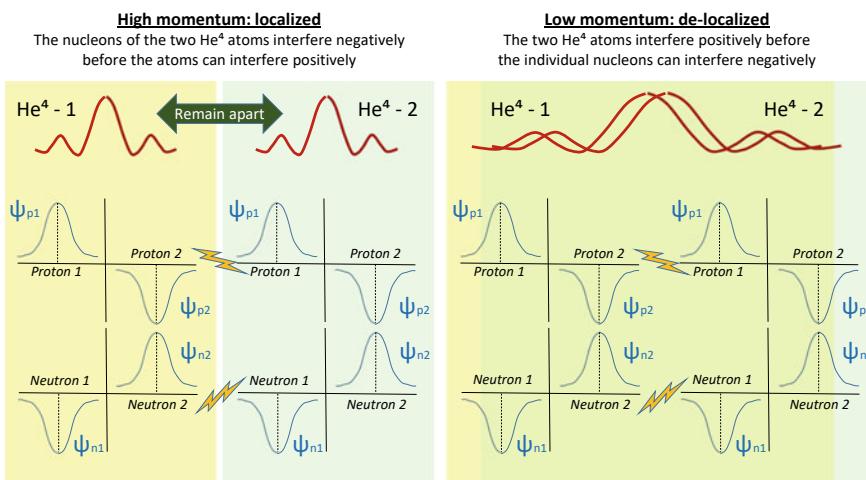


Fig. 13.30 De-localisation as a precondition for He^4 boson behaviour

13.6 Conclusions on Spin, Helicity, and Chirality

Spin can be viewed as a wave going around on a circle. This spin wave is part of the waves that make up a “particle”. That wave can be viewed as a helix. The spin property of waves originates from the relativistic properties of space–time. One could say that the vacuum imposes a “standard” type of circle on all particle waves.

There are two types of spin. The spin wave of a boson goes around on a regular circle and is quantized into integer units, as the waves would interfere positively with integer values of their wavelength. The spin wave of a fermion goes around a Möbius strip and takes two rounds to get back to its original position. The Möbius way of twisting is called a spinor. Hence, the field a fermion waves around in is called a spinor field. As a consequence, a fermion spin wave must be quantized into integer $\pm\frac{1}{2}$ units. This behaviour leads to bosons getting into symmetric states and fermions into anti-symmetric states. This in turn makes bosons most likely to get into the same states, while fermions cannot get into the same states.

The behaviour of fermions is more complex as a consequence of their spin waves going around a Möbius strip. Fermions have a chirality, which is determined by the twist of the Möbius strip. Twisting one way is called left-handed chirality, while twisting the other way is called right-handed chirality. The chirality shows a kind of precession at a frequency determined by the particle mass. Fermions also exhibit helicity. Massive fermions can have negative helicity when the spin direction is opposite to the direction of motion and positive when it is equal to the direction of motion. Positive helicity is associated with right-handed chirality at (almost) the speed of light. Negative helicity is associated with left-handed chirality at that velocity. This is due to the direction the field twists along the Möbius strip.

Remember that these are just ways to get an understanding of spin and how it behaves. What spin really is remains unknown. Many physicists just say that spin is an internal degree of freedom, thereby stating that it does not make sense to explain qualitatively what spin is because we really do not know exactly. We just know its properties in a mathematical sense. I disagree to the extent that it is not only great to get at least *some* feeling of how it might all work, but also because guessing the next step for research may well benefit from such a picture.

Chapter 14

Conservation of Charge and Particle Number



Previously we discussed the Noether current to show how energy and momentum is conserved as a consequence of the fact that laws of physics (that govern energy and momentum) do not change when we translate to a different section of space–time. The Noether theorem states [Ref. 66]:

If a system has a continuous symmetry property, then there are corresponding quantities whose values are conserved in time.

The propagation of these quantities in time is called a Noether current.

Now that we have a good picture of what “particles” are, how they consist of waves, and what electric charge is, let’s revisit the Noether current and see what other symmetries we can find and what conservation laws they lead to.

14.1 Particle Number Conservation

We have seen how fermions are made of a bunch of waves. All these waves represent an energy and momentum density. Let’s look at a number of fermions in a box. There is no contact with the world outside the box so it is a closed system. When we add up all the waves of all the fermions, we get the total energy and momentum of the fermions in the box. Since the energy and momentum are governed by the laws of physics, and these laws are symmetric under translations in space–time, we can say that this total energy and momentum will not change unless impacted from outside.

In addition, there are laws that govern the way fermions can be created or annihilated. A fermion can be annihilated only when it meets its anti-particle. The result is energy, e.g., in the form of a photon. So, energy is conserved in the process. Similarly, a fermion can only be created together with its anti-particle. These laws are also symmetric in space–time. So, we have a system (a box) governed by laws that all are symmetric with respect to translation in space and time.

Consequently, using the Noether theorem, we can say that the number of particles minus the number of anti-particles must be conserved! All this is logical: when there

is no other means to create a fermion except by creating a fermion–anti-fermion pair, the number of fermions minus the number of anti-fermions cannot change. As soon as we add a fermion, we also add an anti-fermion and the number of fermions minus the number of anti-fermions remains the same.

This number cannot change as a whole, but within the system it can move from A to B. Hence, the evolution in time of the number of fermions minus the number of anti-fermions is called the Noether particle current.

You may wonder why there is no other way to create a fermion. After all, it is just an excitation of the particular fermion field. So why could we not excite the field in some other way? In general, there are several ways to excite a particular fermion field. However, the excitation process always requires a pair to be produced. *We cannot excite the field in one direction of time without also exciting it in the other direction of time* (see Sect. 10.1). We have seen that the anti-particle is the same as the particle, except that it moves in the other direction of time (or better, it is the CPT mirror image of the particle, see Sect. 17.5). So, symmetry in time (or rather CPT) is the reason why any fermion cannot be produced without its anti-version.

This raises an interesting question about our universe. We saw before that there is more matter than anti-matter in the universe. Well, if the laws of physics are the same with respect to the direction of time, how did the surplus of fermions get produced? Or, to put it differently, where did all the anti-fermions go? We can fantasize about an anti-matter universe in the other direction of time (see Sect. 10.8) but this is not even scientifically hypothesized, let alone proven. In Sect. 17.5, where we discuss parity violation and CPT symmetry, we will get back to this question from a more scientific point of view. But, as we will see, it is not yet quite resolved!

14.2 Charge Conservation

We have seen that the universe is locally symmetric with respect to phase shifts. The fact that it is locally symmetric gives rise to phase shifts being propagated by a gauge wave (a photon). However, it is still a symmetry! The “power” of the phase shift is determined by the strength with which a fermion connects to the gauge field. That strength was referred to as the charge of the fermion. The charge is simply a measure of how the fermion produces phase shifts.

There is nothing that distinguishes one fermion from another (of the same type!) in terms of how it can produce phase shifts. So, each fermion of one particular type will do it in exactly the same way. Hence, each fermion of the same type has the same charge. That also means that the charge of a particular fermion cannot change in time or space. So again, we have a symmetry: the charge of each fermion remains the same at each position in space–time. Put differently, a fermion can only produce phase shifts in one way and this does not change in space–time.

Charge can only be annihilated or created by annihilating or creating a fermion. In such a process, an anti-fermion is also annihilated or created. The anti-fermion must

have the opposite charge as a consequence of its moving in the opposite direction in time.

Consequently, we can say that, just as fermion number minus anti-fermion number is conserved, so charge is conserved as well. Note that the charge of a fermion cancels against that of the anti-fermion since it has an opposite sign, so we do not have to explicitly state “charge minus anti-charge” is conserved. It suffices to say “charge is conserved”. For example, suppose we have 6 electrons in the box. Then the total charge is -6 . When we produce 4 electrons +4 positrons, the total charge is $-6 - 4 + 4 = -6$. Within the system, the charge may move from A to B. In that case we have a Noether charge current.

In general, we can say that a closed system can have a number of ways (laws) to change a quantity. When all these processes that change the quantity are symmetric in the same way, the quantity is conserved. That means that we have to know two things about the system in order to be able to decide whether a quantity is conserved:

1. We have to know all processes that can change that quantity
2. We have to make sure that all these processes obey the same symmetry.

When these conditions are met, conservation simply means that the bookkeeping is right and there are no *unknown* processes that could change that property. Such an unknown process might change the particle number, charge, energy, etc., and leave us baffled. But once we find out what that process is and we can confirm that it obeys the same symmetry, we can incorporate it in our description of the system and the bookkeeping will work again. So, as we saw before, there is nothing mysterious about conservation laws. When all processes that govern a system obey the same symmetry, conservation is just dull bookkeeping.

Since things can move freely from one place to another and all things move through time, the symmetry of space and time with respect to these processes is relevant. If we found that there was a process that changed the charge of a fermion in a different region in space but did not exist here, space would not be symmetric. Consequently, charge wouldn't be conserved and all it would take to change the charge in the system would be simply to move up and down between the two regions of space.

Chapter 15

Particle Zoo



So far, we have discussed only a fermion wave and the electromagnetic gauge wave. But there are lots of particles. Each particle is an excitation of a different type of field. When we set out to describe these particles, we have to distinguish the associated fields. That means that we have to consider a separate wave, mass potential, and interaction potential for each different field.

15.1 A Visit to the Particle Zoo

Before we can think of describing waves, let's first take a look at some of the particles that have been discovered. The main particles are listed in Table 15.1, including the time of discovery and who discovered them.

Table 15.1 shows some of the particle discoveries made over the last hundred years or so. The left-hand column gives the name and the (Greek) letter for the particle. An anti-particle is indicated by a bar over the particle letter. In general, the anti-particle is then called “particle-bar”, e.g., the proton is called “p” and the anti-particle of a proton is called “p-bar”.

The third column gives the date of discovery. A lot of particles were discovered in the late 1940s and 50s. At that time, the quark had not yet been hypothesized. That happened in the mid-1960s. So, it is easy to imagine the headache all these new particles gave physicists in those days. What did all those particles mean? Where did they all come from? But then, in the 1960s, it was discovered that many of these particles are actually made up of smaller constituents, which were called quarks. The inventor of this term, M. Gell-Mann, wrote an amusing anecdote about the origin of the word “quark”, in which he explained that it came from the phrase “three quarks for muster mark”, taken from “Finnegan’s Wake”, by James Joyce [Ref. 27].

At first only three quarks were known, but later others were discovered. There are in fact 6 flavours of quarks. In the meantime, the carriers of different forces were also discovered. These were the W- and Z-bosons (weak force) and the gluons

Table 15.1 List of some of the particle discoveries in the last 100+ years

Particle	Type	Year of discovery	By	How
Electron e^-	Lepton	1897	J. J. Thomson	Cathode ray tube
Proton P	Baryon uud	1911–19	E. Rutherford	Alpha scattering from nuclei
Photon Γ	Gauge boson	1923	A. Compton	X-rays scattered by atomic electrons
Neutron N	Baryon udd	1932	J. Chadwick	Beryllium bombarded with alpha particles
Positron (anti-electron) e^+	Lepton	1932	C. Anderson	Cosmic radiation
(Anti-) Muon μ^+, μ^-	Lepton	1937	C. Anderson and S. Neddermeyer	Cosmic radiation
Pion π^+, π^-	Meson u <u>\bar{d}</u> , d <u>\bar{u}</u>	1947	C. Powell and group	Cosmic radiation
Kaon K^o	Meson d <u>\bar{s}</u> , s <u>\bar{d}</u>	1947	G. Rochester and C. Butler	Cosmic radiation
Kaon K^+, K^-	Meson u <u>\bar{s}</u> , s <u>\bar{u}</u>	1947	G. Rochester and C. Butler	Cosmic radiation
Pion π^o	Meson u <u>\bar{u}</u> , d <u>\bar{d}</u>	1949	R. Bjorklund and team	Proton accelerator (cyclotron)
Lambda Λ^o	Baryonuds	1951	C. Butler and group	Cosmic radiation

(continued)

Table 15.1 (continued)

Particle	Type	Year of discovery	By	How
Ξ_1	Baryon dss	1952	R. Armenteros and team	Cosmic radiation
Ξ^-				
Sigma	Baryon dds	1953	W. Fowler and team	Accelerator—Kaon beam
Σ^-				
Sigma	Baryon uus	1953	G. Tomasini and Milan-Genoa team	Cosmic radiation
Σ^+				
Anti-proton	Baryon $u\bar{u}\bar{d}$	1955	E. Segre and team	Accelerator—proton interactions
p^-				
Anti-neutron	Baryon $\bar{d}\bar{d}\bar{u}$	1956	B. Cork and team	Accelerator—proton interactions
\bar{n}				
Sigma	Baryon uds	1956	R. Plano and team	Accelerator—Kaon beam
Σ°				
Electron neutrino	Lepton	1956	C. Cowan and F. Reines	Nuclear reactor
$\nu(e)$				
Anti-Lambda	Baryon $\bar{u}\bar{d}\bar{s}$	1958	D. Prowse and M. Baldo-Ceolin	
$\bar{\Lambda}$				
Ξ_1	Baryon uss	1959	L. Alvaraz and team	Accelerator—Kaon beam
Ξ°				
Muon neutrino	Lepton	1962	L. Lederman, M. Schwartz, J. Steinberger	Accelerator—decay from pions
$\nu(\mu)$				
Omega	Baryon	1964	V. Barnes and team	Accelerator—Kaon beam
Ω^-				

(continued)

Table 15.1 (continued)

Particle	Type	Year of discovery	By	How
(Anti-) Up (\bar{u}), u	Quark	1964–72	Gell-Mann and Zweig (theory)	Electron scattering/neutrino scattering
(Anti-) Down (\bar{d}), d	Quark	1964–72	Gell-Mann and Zweig (theory)	Electron scattering/neutrino scattering
(Anti-) Strange (\bar{s}), s	Quark	1964–72	Gell-Mann and Zweig (theory)	Electron scattering/neutrino scattering
J/Psi J/ ψ	Meson $c\bar{c}$	1974	B. Richter and team/S. Ting and team	Accelerator— e^-/e^+ annihilation
(Anti-) Charm (\bar{c})c	Quark	1974	B. Richter and team/S. Ting and team	Inferred from J/ ψ
(Anti-) Tau τ^+, τ^-	Lepton	1975	M. Perl's team	Accelerator— e^-/e^+ annihilation
Charmed Lambda Λ_c	Baryon udc	1975	N. Samios and team	Accelerator—neutrino beam interactions
D°, D ⁺	Meson cu, c \bar{d}	1976	G. Goldhaber and team	Accelerator— e^-/e^+ annihilation
Upsilon Y	Meson $b\bar{b}$	1977	L. Lederman and team	Accelerator—proton interaction
(Anti-) Bottom \bar{b} , b	Quark	1977	L. Lederman and team	Inferred from Upsilon
Gluon	Gauge boson	1979	TASSO and other experiments at DESY	Accelerator— e^-/e^+ annihilation
B°, B ⁻	Meson	1983	Cleo team	Accelerator— e^-/e^+ annihilation
W ⁺ , W ⁻	Gauge boson	1983	UA1 and UA2 teams CERN	Accelerator p/p ⁻ annihilation

(continued)

Table 15.1 (continued)

Particle	Type	Year of discovery	By	How
Z	Gauge boson	1983	UA1 and UA2 teams CERN	Accelerator p \bar{p} collisions
Lambda-B Λ_b	Baryon udb	1991	R422 team CERN ISR	Accelerator p \bar{p} collisions
(Anti-) Top \bar{t}, t	Quark	1995	CDF and D0 teams Fermilab	Inferred from particle decay into W and b
Tau neutrino $\nu(\tau)$	Lepton	2000	DONUT team Fermilab	Neutrino beam
Higgs H	Boson	2012	ATLAS and CMS experiments CERN	Accelerator p \bar{p} collisions

(strong force). The latest famous discovery was the Higgs boson, which gives many of these particle's their mass [Ref. 15]. So, the quarks may have given *some* structure in the particle zoo, but there is much more to tell. Finally, the relation between all these particles has been described by the “standard model”. This model describes all fundamental particles (quarks, leptons, bosons) as excitations in different types of fields.

The best way to discover how they are related is by considering the related forces. Hence, the next two chapters will be about the weak force. This will tell us a lot about the relations between the fundamental particles as well as about the role of Higgs. In the chapter after that we will investigate the strong force. This is responsible for holding quarks together to form protons, neutrons, mesons, and the like. As a secondary effect it is also responsible for tying protons and neutrons into the nucleus of an atom. By then we will have discussed the entire standard model.

The fourth column of Table 15.1 names the people, teams, or facilities that discovered the particles concerned, while the rightmost column shows the way they were discovered. In the 1940s and 50s, a lot of particles were discovered by looking at cosmic radiation. This is radiation that originates from high energy cosmic processes or events, such as supernovae or active cores of certain galaxies. It consists of very high energy photons (gamma rays), protons, and electrons. They hit the outskirts of our atmosphere at extremely high speeds and collide with atoms. In these collisions, a lot of particles are produced from the high energy of the collision. The latter collide with other atoms in the atmosphere, producing even more particles, until finally a whole shower of particles reaches the lower regions of the atmosphere or even ground level. By detecting those particles, we can learn what the original particles were. In those days many different types of particles have been found in this way.

Later, we learned how to accelerate particles and make them collide with a target, or with each-other. If the colliding particles are each-other’s anti-particles, their mass energy would add to the collision energy in producing new particles. The most powerful accelerator operating this way is the Large Hadron Collider (LHC) at CERN in Geneva [Ref. 70]. It smashes protons and anti-protons onto each other at very high energies. The advantage of such machines over cosmic rays is that we know, not only what particles went into the collision, but also their speed (energy) and direction. Another advantage is that we can detect the collision from the start. This all makes it easier to interpret the collision and its products, including the exotic ones just after the collision.

The second column of Table 15.1 shows how each particle is classified. When a particle consists of quarks, as all mesons and baryons do, their quark content is shown by a letter code. Each letter represents a quark flavour. For example, the proton consists of uud, which means an up quark (u), another up quark (u), and a down quark (d). There are several ways to classify particles. One way we have seen before: fermions versus bosons, based on their spin (see Fig. 15.1). However, there is a subdivision. Fermions fall into two categories: leptons and hadrons. The leptons are fundamental (or elementary) particles, in the sense that so far we do not know of any substructure in these particles. The hadrons on the other hand, are composite particles: they are made of quarks. To complicate things, not all hadrons

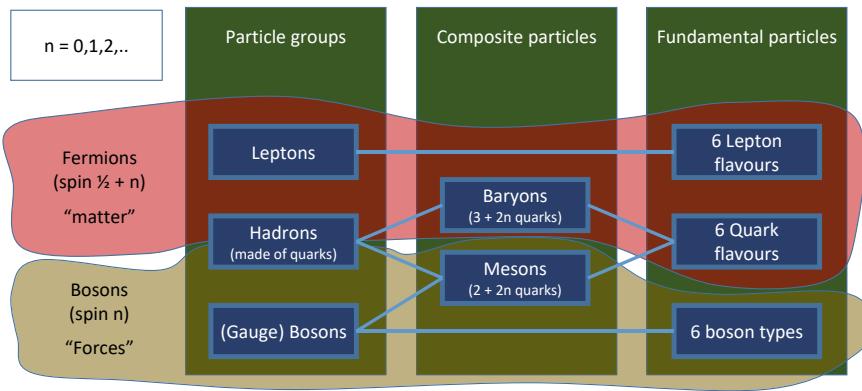


Fig. 15.1 Types of particle

are fermions. Only those that are made of three quarks are fermions. They get the subcategory “baryons”. The other subcategory of hadrons are the “mesons”. They are made of two quarks and are in fact bosons, since they have spin 0 or 1. Apart from leptons and hadrons, there are the gauge bosons. So some bosons are fundamental and others, the mesons, are composite. See Fig. 15.1 for an overview of these groups and how they are related.

The fundamental leptons separate into 6 lepton flavours. First, there are the electron, muon, and tau, which exhibit the same behaviour, charge, etc. They only differ in their mass. Then, each of these three particles has an “accompanying” neutrino. We will see later how these are related. The three leptons and their neutrino counterparts are often referred to as three families of particles, although they are also sometimes called generations. However, I like families better.

The building blocks of the hadrons, the quarks, are fundamental as well (as far as we know). There are six quark flavours. They too pair up into three pairs that can be organized into three families. The pairs of the different families look alike, but again they differ in mass.

When we rank them according to mass, we get three families with each family containing two leptons and two quarks. The family with the lightest leptons and quarks consists of the electron, its neutrino friend, and the up and down quarks. The up and down quarks are the constituents of the proton and neutron, as well as pions. Therefore, this family makes up the ordinary matter we find around us in everyday life.

The last group of fundamental particles are the bosons. There are six types, but these do not relate to the families just discussed. There is the photon, which is the gauge wave of the electromagnetic interaction. Then there are three bosons propagating the weak force. We will discuss those later, in the chapter about the weak force. There is the gluon, which exists in combinations of three different “colours” and “anti-colours”. There turn out to be 8 different combinations. This will also be treated later, in the chapter about QCD. The final fundamental boson is the Higgs

boson, which is responsible for giving mass to particles. This, too, will be discussed in the chapter about the weak force.

In the last chapter we discussed the chirality of particles. We saw that the left- and right-handed particles should be considered as different types of particle. All fundamental leptons and hadrons exist in a left-handed form. Except for the neutrino, all particles also have a right-handed “counterpart”. We also met anti-particles before. So, each fundamental particle as well as its right-handed counterpart (when it exists) has an anti-particle version.

15.2 Introducing the Fundamental Particle Overview

All these particles are summarized in Fig. 15.2. They are categorized into the three families.

The overview in Fig. 15.2 suggests some structure, but so far, we have only described the U(1) symmetry, represented by the grey area. This symmetry generates phase shifts, which we identified as the origin of the electromagnetic force. All fermions except neutrinos are charged and hence generate and feel phase shifts. It turns out that more symmetries exist between the particles. These symmetries generate other types of force. Such symmetries will show up when we start to look at the individual particle waves instead of the generic group of “fermions”. We will see symmetries appear that have different properties than the phase shift we discussed.

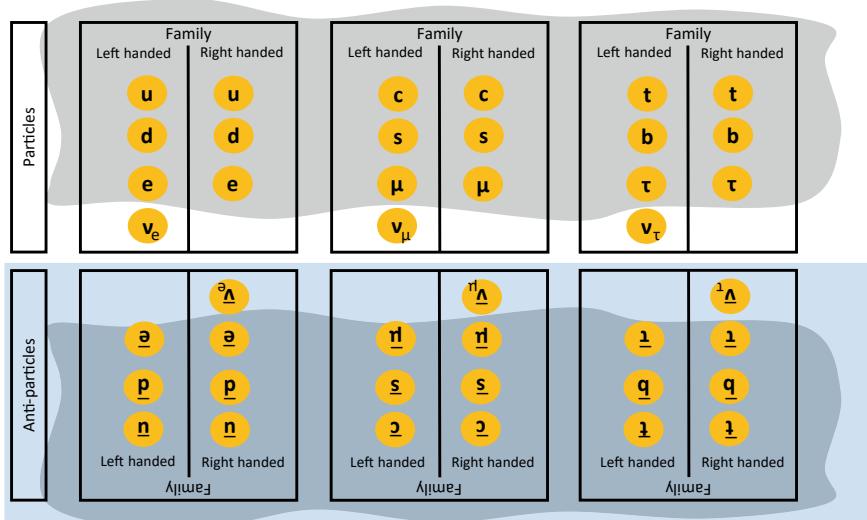


Fig. 15.2 Overview of fundamental fermions. The grey area represents the U(1) symmetry that they are all subject to. Only neutrinos have zero charge, but otherwise all fundamental fermions are charged, meaning that they generate and feel phase shifts

However, the broad idea is the same: each symmetry will have to be made a local symmetry due to the maximum velocity in the vacuum. This will lead to gauge waves, similar to the photon gauge wave that is connected to phase shifts. Such gauge waves exist for the weak force, which is a consequence of an SU(2) symmetry. These gauge waves become the W- and Z-bosons. Gauge waves also exist for the strong force, which is a consequence of an SU(3) symmetry. These gauge waves become the gluons. One thing stays the same: this whole complex world of particles is in fact just a world of waves. Each different particle is an excitation (a wave) in a different type of field and produces different types of symmetric shifts and turns, leading to a variety of gauge waves.

So, remember, any time we are talking about particles, we are in fact talking about the idealized excitation of the field that goes with that particle. In reality, each such excitation consists of a cloud of disturbances in different fields, just as the electron turned out to be a cloud of disturbances. This means that many particles are in fact mixes of different types of particles, as we will see.

15.3 The Rabbit Hole

If you think you just got anything like a complete overview, get ready for a disappointment. The rabbit hole goes a lot deeper. The table of particles is far from complete. There are other mesons, baryons, and exotic versions of those, many of which have been found quite recently. An example of exotic quark combinations are tetraquarks and pentaquarks. The former are “mesons” that are made of two quarks and two anti-quarks. So, in fact they look like a double meson. It is assumed that they are made of two mesons that are weakly bound. However, they decay via the strong interaction into their constituent parts, which are normal mesons. Pentaquarks are made of five quarks: four quarks and one anti-quark, so they look like baryons with two extra quarks. These exotic types are thought to be made of a baryon and a meson that are weakly bound together. Indeed, they decay via the strong interaction into a baryon and a meson. For instance, CERN reported in 2015 that they had found some evidence for the existence of pentaquarks [Ref. 26]. In 2019 a new pentaquark was discovered, strengthening the evidence for such five-quark states. The particle, named $P_c(4312)^+$, decays to a proton and a J/ψ particle (composed of a charm quark and an anticharm quark) [Ref. 88]. In 2020 two tetraquarks were discovered ($c\bar{s}\bar{u}\bar{d}$, composed of D mesons, and $cc\bar{c}\bar{c}$, composed of two J/ψ mesons) [Ref. 87]. In 2021 CERN reported that they had found a slightly less unstable tetraquark, made of $cc\bar{u}\bar{d}$.

One interesting issue is the mass of such tetraquarks. As we saw before (see Sect. 6.2), a bound state has a lower mass than the sum of the masses of the constituent particles. So, it would take energy comparable to the “missing” mass to break that state. Predictions of the mass of the $cc\bar{u}\bar{d}$ state varied substantially, from 250 MeV below to 200 MeV above the $D^* + D0$ mass [Ref. 87], i.e., the mass of its constituent mesons. Its observation at CERN reveals that it is a mere 273 ± 61 keV below the mass of the two D mesons [Ref. 87]. So, it should be a bound state! However, due

to the uncertainty relation it can temporarily have enough energy to decay. This mechanism, though, is a slow decay mechanism, much slower than if the mass of the tetraquark were in the neighbourhood (or above) the mass of its constituents. The fact that the mass is below that threshold means that there is a significant binding energy. That means that there is a clear mechanism for the binding of mesons in a tetraquark. Such a mechanism may be the exchange of light mesons, comparable to what ties protons and neutrons together in an atomic nucleus (see also the residual strong force or “nuclear force” in Sect. 18.4). This raises the expectation that an even heavier tetraquark $bb\bar{u}\bar{d}$ might be stable with respect to the strong interaction [Ref. 87], which would mean that it does not decay into regular mesons, or would take a very long time to do so.

There are also resonances. These are mesons or baryons that have a very short lifespan, on the order of just 10^{-24} s before they decay into something else. An example is the delta particle, which is a resonance of the proton when it is bombarded with pions of a certain energy. Basically, what happens is that the pion annihilates one of the quarks in the proton and replaces it with another. For example, a π^+ consists of an anti-down quark and an up quark. The proton has two up quarks and a down quark. The down quark in the proton annihilates against the anti-down in the π^+ . At the same time the u in the π^+ becomes part of the “proton”. The “proton” now consists of three up quarks. This is one of the existing delta particles. It takes on the order of 10^{-24} s before the delta falls back to a uud state (a proton), while emitting a π^+ .

This process can be compared with the absorption of a photon by an atom. The atom gets into a higher state of energy and falls back to its ground state while emitting a photon again. The short existence of a delta particle can be viewed as an excited state of the proton. The delta is then called a resonance. Such a resonance cannot be measured directly. It is usually measured by an absorption peak in the pion beam bombarding the protons. The width of the peak is an indication of the lifespan of the resonance. The shorter the lifespan, the more uncertain the energy of the process. Hence, the wider the energy peak of the resonance. When pions have exactly the resonance energy, they get absorbed. But when the lifespan of the resonance is shorter, the pions will also get absorbed when their energy is slightly off, because of the uncertainty in energy of the resonance. As you can imagine, this goes for many other mesons and baryons, leading to many different types of resonances.

And now the good news: so far, all these (exotic) particles and resonances are still made of the fundamental particles shown in Fig. 15.2. Hence, we will use some of the particles in the overview as an example, but we will focus on the fundamentals. Put differently, we will discuss the different types of Lego® blocks, but that does not mean that we will have to discuss the endless number of things you can make with those blocks.

So, in Chaps. 16 and 17 we will start to take the individual particle waves of the electron and neutrino (instead of a generic fermion) into consideration and we will include the SU(2) symmetry. We will discuss where that leads us and we will extend this to some of the quark behaviour. In Chap. 18 we will go into this more deeply

and add the SU(3) symmetry to see what waves we get out of that. Finally, we will find our way out of the rabbit hole and complete the picture.

Chapter 16

Electroweak Force in the Early Universe



If we look at the zoo of particles and compare them, they suggest a certain structure. One thing that stands out is that some particles have masses that do not differ much. If masses were randomly determined for particles, the chances of two particles having such close masses would not be great so something special must be going on. An example is the proton and the neutron. Their masses are:

- Proton: $1.672621898(21) \times 10^{-27}$ kg
- Neutron: $1.674927471(21) \times 10^{-27}$ kg.

So, this is very close indeed. One might wonder whether they are just two faces of the same particle? That is an interesting idea. What does it mean for a particle to have two faces? It would mean that the particle is able to change some of its characteristics by “rotating” in “some other dimension” (see Fig. 9.3 and Sect. 9.1). When two particles are really different, they cannot “turn” into each other. Such a change would not go without a ripple in the vacuum. It can be compared to a phase shift and will propagate some potential away. Hence, we suspect that some type of force must go with such a change. The fact that a particle has two faces and can turn one face into the other is a type of symmetry.

For the proton and the neutron this would mean that they primarily change charge when they turn into one another. The problem with the proton and neutron is that their mass is not exactly the same and they are composite particles. They consist of up quarks and down quarks. So, might these two types of quarks be one particle with two faces? They turn out to differ much more in bare mass and in charge. So, is all this a coincidence after all?

Not really, but to understand that, we need to go back in time. Way back in time. In fact, back to the first fraction of a second of the universe. As it turns out, some of the symmetries in the universe do exist, but only above a certain energy level. A level that existed only at the very beginning of the universe. Below that energy these symmetries got broken and what we see in our current universe are the remnants of them.

The breaking of symmetry has an effect on the U(1) symmetry we discussed as well. Since this is a complicated matter, we start by going back to the beginning of the universe, when life was still simple and symmetries were recognizable as a symmetry, and we pick up from there.

16.1 The First 10^{-12} s

The current theory about the universe is that it started with a “big bang” some 13.8 billion years ago. This big bang theory describes how the universe expanded from a very small size, when it had a very high density and temperature, to the current universe we live in. This theory can explain a number of phenomena we see in the current universe and when we look back into history. One of these phenomena is the observation that all galaxies move away from each other, and also that the further apart they are, the faster they move away. The best way to explain that is by an expanding universe.

Another observation is the 2.7 K background radiation. This radiation is extremely homogeneous in all directions throughout the universe. How can that be? Such homogeneity in temperature would require that this radiation was in thermal equilibrium at some point in time. That in turn requires the radiation to have been in contact so that it can equalize temperature differences, something that is not possible when that radiation is billions of light years apart. The big bang model can explain this by assuming that the radiation was in fact in contact at a very early stage of the universe when it was still very small, at a time when the universe had a temperature of 3000 K. This was the temperature below which electrons got bound with protons to form hydrogen atoms. Before that, light would be scattered, absorbed, and emitted continuously by all the loose electrons and protons, and the universe would have been one big fog. After that, light would no longer be scattered or absorbed all the time, and the universe would have become transparent to radiation. This radiation has cooled down since then, as the universe expanded. You could say that the wavelength of that radiation stretched together with the expansion until it became radiation that agrees with a temperature of 2.7 K. The current structure in the universe originates from slight differences in the homogeneity of the mass-energy distribution at the beginning of the universe. Those slight differences must have been there when the universe became transparent, leading to small temperature differences in the radiation at that time. These should be visible in the otherwise homogeneous background radiation. Such differences were measured in the background radiation by the Cobe satellite [Ref. 25], providing another confirmation of the big bang theory.

By considering the velocities and distances of galaxies, we can calculate back when these should have been concentrated in one point. That calculation gives us the age of the universe as 13.8 billion years.

The theory also explains the large proportion of light elements in the universe. Basically, the processes in the early universe allowed for condensation into atomic structures as of the moment the universe cooled below 3000 K. The first elements

that could be formed were hydrogen and some helium, the two lightest elements. The other elements were all produced in stars that formed later. The processes in stars that are responsible for creating the elements are well understood and can explain the relative abundances of lighter and heavier elements in the universe.

Hence, there is good evidence pointing in the direction of the big bang theory, but there are also unsolved problems. In this book we will not go into this. We will be using some of the relevant outcomes of this theory. The current theory does not describe how it all started, but it can more or less accurately describe what happens as of the first 10^{-43} s. Almost time zero, but not quite! The temperature of the entire universe at that stage would have been a staggering 10^{28} K. In these first fractions of a second lots of things would have happened. The universe would have expanded very fast and the temperature would have dropped quickly. At 10^{-12} s it would have dropped to 10^{16} K. Still an extremely high temperature, but it took only a fraction of a second to get there.

It is at this temperature that the electroweak symmetry would have been broken. So, what is electroweak symmetry? And what exactly happens when a symmetry breaks?

16.1.1 Electron and Neutrino Waves

The major difference with our current universe is that, before symmetry breaking, the Higgs field was 0. After the symmetry broke, the Higgs field was homogeneously present throughout the universe. This has a number of consequences. One of them is that the mass of all fermions was 0 before symmetry breaking. We said before that fermion masses originate from their interaction with the Higgs field. So, when the Higgs field was non-existent (or rather, when its springs had zero strength), all fermions must have been massless.

Since the up quark and down quark have no mass at all at this stage, the difference between them has gone, and that suggests that they might be two faces of the same particle. The same goes for the electron and the neutrino. Both are massless before symmetry breaking. So, let's look at this a bit more closely. We take the electron and the neutrino as separate fields leading to separate waves. This means that instead of a single fermion wave, we would describe two waves: a neutrino wave and an electron wave. Another consequence is that we would have a mass potential for both. However, their mass is 0 before symmetry breaking and so there is no mass potential. Or better, there is a mass potential, but it contains the coupling strength with Higgs and Higgs is 0.

Furthermore, we have to make a distinction between left-handed and right-handed waves. In the section where we discussed this difference between left-handed and right-handed, we considered them as different particles. We saw that their interaction with Higgs can make them switch from one to the other. We will discuss this later. For now, the Higgs field is 0 and the left-handed and right-handed waves *cannot* switch from one to the other. Hence, we have to consider them as two different types

Table 16.1 Three fermion wave types that represent three different particles

Field	Wave type	Mass “spring”
Right-handed electron	Fermion	0
Left-handed electron	Fermion	0
Neutrino	Fermion	0

of particle. They simply are not the same wave, and as we will see later, they do behave differently.

There is one interesting problem: a neutrino exists only as a left-handed particle. It is not known why there is no right-handed neutrino. This too suggests considering the left- and right-handed particles as different breeds.

Consequently, we have to split the electron wave up into two waves: a left-handed electron wave and a right-handed electron wave. Summarizing, we end up with three wave types (see Table 16.1).

In this overview we have not yet considered the U(1) symmetry that leads to phase shifts and an electromagnetic force.

16.1.2 Introducing the Original U(1) Gauge Field

We previously discussed the U(1) symmetry, which leads to the possibility of phase shifts. This in turn leads to an interaction potential and a gauge wave. The interaction potential contains a coupling constant just as we have seen before. This coupling constant should be the charge of the particles. However, symmetry breaking has an impact on the U(1) symmetry as well. So, before symmetry breaking, the charge is different. This charge is called “weak hypercharge” and is denoted by Y . Both the electron and the neutrino have a weak hypercharge. Remember that the neutrino in our current universe has no charge, so here we have one of the differences in charge before and after symmetry breaking.

Another consequence is that the gauge field that is created is not exactly the same as the electromagnetic field as we know it in our current universe. Therefore, the excitations of this field are not called photons. They are called B-bosons.

The weak hypercharge of the left-handed electron and the neutrino is $Y = -1$. The right-handed electron has a weak hypercharge $Y = -2$. Why is that? To be honest, we cannot measure the weak hypercharge since we cannot create the circumstances from before symmetry breaking in our detectors. The right circumstances imply recreating the situation in the universe before 10^{-12} s. This would mean creating a temperature above 10^{16} K. The most powerful particle accelerator in the world today is the Large Hadron Collider (LHC) at CERN, which can create a temperature of the order of 3×10^{12} K in interactions. This is still many orders of magnitude too small to create the right circumstances.

So, we have to rely on theory. The theoretic approach has been basically “calculating backwards” to get the results for weak hypercharge. Our present universe

Table 16.2 Extending our wave types with the U(1) gauge wave and their mutual interactions

Field	Wave type	Mass “spring”	Interaction “spring”
Right-handed electron	Fermion	0	With B-boson
Left-handed electron	Fermion	0	With B-boson
Neutrino	Fermion	0	With B-boson
B-boson	Gauge boson	No	With charged fermion

looks the way it does, and symmetry breaking must deliver precisely our current universe. So we now take the process of symmetry breaking and put in the condition that it must deliver our present-day universe. The result of that procedure is that we must put in properties such as weak hypercharge with the values mentioned here. We will see a bit more about how this works later.

Let's add the gauge wave and its interactions to the overview of fermion waves in Table 16.1. The result is given in Table 16.2.

16.2 Symmetry Amongst the Waves

When we still had one fermion wave, we looked at the Noether current. This current showed that all the complicated waves and disturbances taken together add up to a number of quanta (fermions) that is conserved (particle number conservation). The corresponding current is the flow of fermions from A to B. So, what would happen when we split that one fermion wave into multiple particle waves? The Noether current we calculate from this still gives a conserved number of quanta, but it conserves only the *total* number of quanta. However, it does not say anything about the individual electrons or neutrinos being conserved! The corresponding current is still only the current of electrons + neutrinos together that flows from A to B. Consequently, this allows electrons and neutrinos to change into each other on the way and the total number of electrons + neutrinos is still conserved. From this we conclude that, when we have multiple fermion fields such as the electron field and the neutrino field, we get the *possibility* of internal symmetries! Electrons *could* change into neutrinos. If they can actually do that, they are different faces of the same particle!

But hold on! Can any wave just turn into another as long as they are both fermions? Well, there are some requirements. It is obvious that when waves do not look like each other, turning one into the other is not possible or requires an interaction that deals with the differences. In a perfect symmetry, it requires the waves to look like each other. Let's see what that means for our electron and neutrino.

Table 16.3 The weak hypercharge of left-handed fermions (left) and right-handed fermions (right)

Left-handed chirality	Weak hypercharge Y_w	Right-handed chirality	Weak hypercharge Y_w
$\gamma_e, \gamma_\mu, \gamma_\tau$ neutrinos	-1	Neutrinos (probably) non-existent	x
e, μ , τ leptons	-1	e, μ , τ leptons	+2
u, c, t quarks	+1/3	u, c, t quarks	+4/3
d, s, b quarks	+1/3	d, s, b quarks	-2/3

First of all, a left-handed wave is not the same as a right-handed wave. They are each other's mirror images, and when there is no interaction that can turn one into the other, they must remain different waves. So left-handed electrons will not turn into right-handed or vice versa, at least not before symmetry breaking.

Then one possibility remains: the left-handed electron and the neutrino (which only comes in left-handed versions). They have the same weak hypercharge and both are massless. So, these waves look a lot like each other. We have a true symmetry here, and nothing stands in the way of them changing into each other. The left-handed neutrino and the left-handed electron are two faces of the same particle!

The same goes for some other particles. For example, the left-handed up quark and down quark are a pair that share the same weak hypercharge. However, their right-handed versions do not! So, unless there is an interaction that can change their weak hypercharge when they turn, they cannot turn into each other. What might be such an interaction? When two particles change from one into the other, they must create a wave. Just as with a phase shift, a change will be created that needs to be propagated away and carries momentum.

Let's look at the weak hypercharge of the fundamental fermions to get an idea of which particles could turn into each other (see Table 16.3).

The up, charm, and top quarks are different particles that belong to different families (see Fig. 15.2). They are unable to turn into each other. The same goes for the electron, muon, and tau. The up and down quark belong to the same family and so they could still turn into each other.

The same goes for the other leptons and quarks. Their left-handed version can turn into their counterpart within their own family:

- Left-handed quarks: u and d, c and s, t and b
- Left-handed leptons e and γ_e , μ and γ_μ , τ and γ_τ .

But none of their right-handed versions can turn into their counterpart. Therefore, we must conclude that right-handed particles are not two faces of the same particle. They are different particles altogether.

All of the particles that can turn into each other are pairs. Each member of the pair is a whole wave function. Unlike the U(1) symmetry for phase shifts, the entire wave now shifts. Since a wave is represented by a phase dimension on the horizontal

axis and an amplitude (field strength) dimension on the vertical axis, we have a two-dimensional object that changes into another two-dimensional object. A symmetry between two two-dimensional objects was described by a rotation in three dimensions, belonging to the group SU(2), as we discussed in the section on symmetry groups.

When the particles are not measured, they cannot be distinguished. Hence, when not measured, they can be any mixture of the two wave functions. When they are measured, we always measure either ψ_1 or ψ_2 , but never the mixture.

So, what are the implications of this symmetry?

16.2.1 Introducing the SU(2) Gauge Field

Just as the U(1) symmetry gave rise to a gauge wave and a coupling constant, we will see that the same goes for the SU(2) symmetry as well. The SU(2) symmetry is a global symmetry. This means that when a particle turns from one face to the other, that change is propagated at infinite velocity throughout the universe. But the vacuum does not allow a change to propagate at infinite velocity. Hence, just as in the U(1) case, we must make SU(2) a local symmetry. The change is then propagated away at light speed (for massless gauge waves) or slower (for massive gauge waves).

So, a change from one particle to another produces a gauge wave that propagates that change away. Just as the U(1) phase shift produces a wave, an SU(2) change produces a wrinkle, but of a different type. So, when a fermion switches to its other face, it produces such a wrinkle. If the wrinkle meets another fermion, it un-wrinkles by turning that fermion into its other face. At the same time, the wrinkle can transport a potential difference just as the U(1) gauge wave does. In fact, mathematically, the SU(2) change is treated very similarly to the U(1) phase change [e.g., Ref. 30, p. 482]. It can be viewed as a phase change that inhibits a wrinkle and that carries a potential difference. This potential difference changes the momentum and energy of both the particle that produced the wrinkle and the particle that absorbed it. Hence, a force is transferred.

We call this gauge wave, this wrinkle, a “W-boson”, similar to the B-boson that propagates a phase shift. When this change is absorbed by another particle, it changes accordingly. So, when an up quark changes into a down quark, a gauge wave (W-boson) is produced that can be absorbed by, e.g., another down quark that will change into an up quark. Figure 16.1 shows a Feynman diagram for this process.

Whether or not a particle will feel the gauge wave depends on the coupling constant. In the case of U(1), this was the (hyper)charge. In the case of the SU(2) symmetry, this coupling constant is called weak isospin (I). Weak isospin is not to be confused with spin. It is something completely different. The only agreement between the two is that both fermion spin and the symmetry between particles are described by the same mathematics of the SU(2) symmetry group.

Since the right-handed particles cannot turn into each other, they have a weak isospin equal to 0. The weak isospin of left-handed particles is $1/2$. With respect to

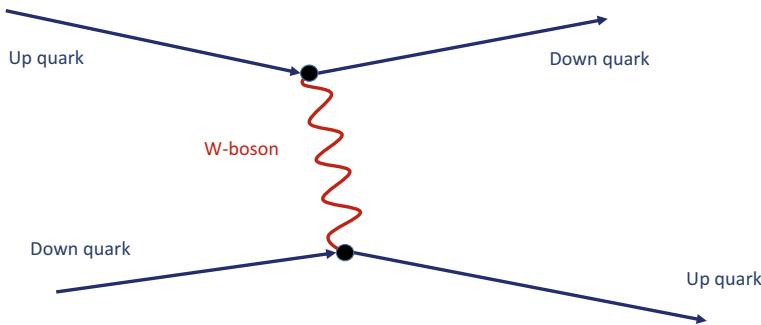


Fig. 16.1 An up quark changing into a down quark produces a W-boson that can be absorbed by another down quark that consequently changes into an up quark

spin, when a particle has a spin of $\frac{1}{2}$ it means that it can be in either of two states: spin $+\frac{1}{2}$ (spin up) and spin $-\frac{1}{2}$ (spin down). The same goes for the isospin I_z . Particles with isospin $\frac{1}{2}$ can be in the state $+\frac{1}{2}$ or the state $-\frac{1}{2}$. These states correspond to the two faces of the particle. So, isospin $-\frac{1}{2}$ is an electron and isospin $+\frac{1}{2}$ is the electron neutrino. Similarly, $+\frac{1}{2}$ is “quark up” and $-\frac{1}{2}$ is “quark down”. We can summarize this for all fermions in Table 16.4.

When we discussed the U(1) symmetry, we were talking about one wave that produces a phase shift. That is a single change. You could view this as the wave changing into itself with a different phase. When a particle changes face, we are confronted with an SU(2) symmetry. In this case, one particle can change into the other or the other can change into the first, but either can also change into itself. This means that there are three different ways of rotating, just as we concluded in the section on symmetry groups. In that section we compared the three ways of rotating as rotations around three different axes in a three-dimensional space. Consequently, we can compare the situation to rotating dice (see Fig. 9.5).

So, we are talking about three different types of rotation. Suppose there is a shoe lying in the water. You can rotate it about three different axes. You can imagine that when you turn it, you create a wave in the water. But when you turn it about one

Table 16.4 The hypercharge and weak isospin states of the fundamental fermions before symmetry breaking

Left-handed chirality	Y_w	Weak isospin states I_z	Right-handed chirality	Y_w	Weak isospin states I_z
$\gamma_e, \gamma_\mu, \gamma_\tau$ neutrinos	-1	$+\frac{1}{2}$	Neutrinos (probably) non-existent	x	0
e, μ , τ leptons	-1	$-\frac{1}{2}$	e, μ , τ leptons	-2	0
u, c, t quarks	$+1/3$	$+\frac{1}{2}$	u, c, t quarks	$+4/3$	0
d, s, b quarks	$+1/3$	$-\frac{1}{2}$	d, s, b quarks	$-2/3$	0

axis (say about an axis parallel to the water) you get a different type of wave than when you turn it about another axis (say about an axis perpendicular to the water). The three possible rotations for a particle also lead to three different types of waves. These three different types translate into three different bosons to carry the weak force. We call these W1, W2, and W3.

$U(1)$ is an Abelian group. This means that it does not matter in what order the two phase shifts are performed. The result is always the same. $SU(2)$ is a non-Abelian group. As we could see from the dice, two different turns give a different result when they are performed in a different order. That is how the symmetry works. So how should we interpret this?

Let's compare two situations:

1. We have a particle 1 that turns into particle 2. With the change, it produces a particular W-boson, say a W1. After that, particle 2 turns into itself. With that change it produces a different W-boson, say a W3.
2. We have a particle 1 that turns into itself. With the change, it produces a W-boson. This is the same boson as the W3 in the previous situation. After that, particle 1 turns into particle 2. With the change it produces the same W-boson as in the previous situation, the W1. This must be so since it is the same particle 1 that rotated.

The first interesting observation is that the two situations end up with the same end result: the particle is in its “particle 2” state, and the same two W-bosons have been produced. However, the end state was reached through two changes in *different* orders. The symmetry requires that the end result *cannot* be the same, just as for the dice!

So how can the end result not be the same? After all, particle 2 is the same in both cases. So, what is different? The only possibility is that the W-bosons are somehow different in the end state. It might seem that this requires an extra degree of freedom in the W-bosons. With that degree of freedom, the W-bosons might be able to memorize the fact that there were previous rotations. However, that seems a bit strange. Another option is that an equal end state actually *is* possible, although it would require extra rotations to get this done. Just look at the dice in Fig. 9.5. The two end states can be made equal by a further rotation of each dice. The only way to do this is to allow the W-bosons themselves to make an extra rotation! But if they made an extra rotation, new W's would be produced. The only way is for them to keep the extra rotation internally. Put differently, they propagate away the extra rotation. Being able to rotate by themselves means that they must carry isospin charge themselves!

So, we can conclude that the $SU(2)$ symmetry demands that the three W-bosons carry isospin charge themselves. Hence, the W-bosons couple to their own field. Unlike $U(1)$ where the photon does not carry charge and cannot produce phase shifts itself, the W-bosons can produce rotations by themselves. This means that a W-boson can create a (virtual) W-boson. We have seen that an electron and neutrino have isospins $+1/2$ and $-1/2$, respectively. The three rotations they can make are carried by three different W-bosons. Now that we see that W-bosons can create the same rotations, what isospin should they have? For the W-bosons this works differently.

Table 16.5 The hypercharge and isospin states of some fundamental bosons before symmetry breaking

Boson	Hypercharge Y_w	Isospin I	Isospin states I_z
B	0	0	0
W1	0	1	+1
W2	0	1	-1
W3	0	1	0

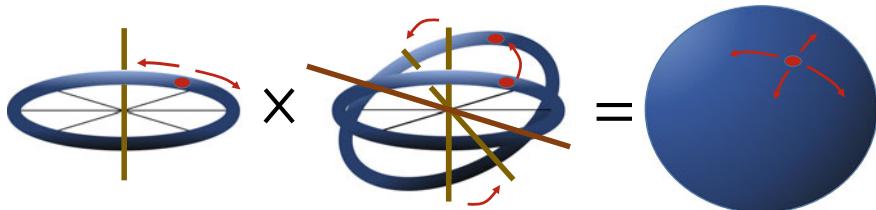


Fig. 16.2 A dot undergoing two symmetry operations: it can reach the product of the two symmetries, which is a sphere

The electron can only create a rotation towards a neutrino and back to itself. But a W-boson can create all three rotations. Therefore, a W-boson gets an isospin of 1, with three state values $I_z = -1, 0, +1$. Each state I_z corresponds to one type of W-boson.

So now we are dealing with two types of symmetry: U(1) and SU(2). They lead to the weak hypercharge Y , carried by the B-boson, and the weak isospin, carried by the three W-bosons. So we can now summarize the hypercharges and weak isospin states of some of the fundamental bosons in Table 16.5.

We now have two symmetries and particles obey both of them at the same time. How does that work? Suppose we have a dot on a wheel, then as the wheel turns, the dot traces out a circle. It can reach any position round the wheel. But now suppose the wheel itself can turn around an axis along one of its diameters (see Fig. 16.2). In that case, the dot gets moved around by two symmetry operations at the same time: going around the wheel and the wheel itself turning around an axis in its plane.

Consequently, the dot can now reach any position on an entire sphere. We call that sphere the product of the two symmetries:

$$\begin{aligned} \text{Spherical symmetry} &= \text{"Going around the wheel"} \\ &\times \text{"the wheel turning on an axis in its plane"} \end{aligned}$$

This product is again a symmetry! In this example, a spherical symmetry. Similarly, the total symmetry of particles with both hypercharge and isospin is called a $U(1) \times SU(2)$ symmetry. Let's not make this more complicated than it is. It just means that a particle obeys both symmetries: it can turn into its other SU(2) face as well as produce U(1) phase shifts.

16.2.2 Including Isospin Symmetry in the Overview of Waves

Now let's try to include the Isospin symmetry. This will lead to a lot of new waves and potentials. Each W-boson must be described by a wave, so three new waves enter the overview. Each electron chirality and the neutrino get an interaction potential with each W-boson. This we will summarize as a general Isospin interaction “spring” for each electron chirality and the neutrino. In addition, since SU(2) is non-Abelian there also has to be an extra potential that describes the interaction (isospin) between the W's!

And so, we can extend the overview of fields for the early universe to include the left- and right-handed electron, the electron-type neutrino, a U(1) symmetry, and an SU(2) symmetry (see Table 16.6).

All the interaction springs represent a potential that is propagated away when a rotation/shift occurs. This potential can induce the same rotation/shift in the next particle. U(1) leads to a “normal” interaction. The further away, the less interaction. The more B-bosons there are, the more interaction. This is called a “linear” interaction. It corresponds to the behaviour we are used to from, e.g., the electromagnetic force.

However, the Isospin symmetry does not show this pattern. The reason is the interaction potential between the W's. The fact that they can interact with each other gives the force a non-linear character. For instance, what might happen is that W-bosons produce more W-bosons, which in turn produce more W-bosons. So, at a greater distance, there is a higher probability of producing more and more W-bosons. This way, the force would get stronger at greater distances! This is not the case for W-bosons *after symmetry breaking*, due to their enormous mass (as we will discover), but we do see this type of behaviour from quarks. We will get back to the non-linearity of the force when we deal with the colour field. We will see then that this type of behaviour leads to the so-called asymptotic freedom of quarks.

There is no interaction potential for the right-handed electron with the W's. This is a consequence of the fact that its isospin (coupling strength to the W's) is 0. Hence, the right-handed electron gets transformed trivially under SU(2), i.e., it stays as it is, while the left-handed electron turns into the neutrino and vice versa.

Table 16.6 Some fields and wave types existing in the early universe and their mutual interactions

Field	Wave type	Mass “spring”	Interaction “spring”
Right-handed electron	Fermion	0	With B
Left-handed electron	Fermion	0	With B, W1, W2, W3
Neutrino γ_e	Fermion	0	With B, W1, W2, W3
B-boson	Gauge boson	No	With charged fermions and bosons
W1	Gauge boson	0	With γ_e , e (left), W1, W2, W3
W2	Gauge boson	0	With γ_e , e (left), W1, W2, W3
W3	Gauge boson	0	With γ_e , e (left), W1, W2, W3

The two symmetries give rise to the following Noether currents:

- Isospin current: a current that preserves left-handed particles, but not the individual left-handed electron and the neutrino. Hence, they may turn into each other, as long as the total number of left-handed particles is conserved.
- Hypercharge current: a current that preserves hypercharge. So, the total sum of hypercharge in a system will remain the same. This is logical as the hypercharge of a particle does not change, even when it switches to its other face (both faces have the same hypercharge), and the total number of particles does not change.

16.3 Introducing the Higgs Field

All fields are all pervasive in space. However, in order to connect to it, the field strength must be non-zero. If the field strength is 0, the connection will be without any effect. So, what type of field could be non-zero everywhere? By now we are most familiar with the electromagnetic field. Could such a field work? Suppose the electromagnetic field were non-zero everywhere. Something strange would happen. It would imply that there is a non-zero magnetic field everywhere. But a magnetic field gives a direction in space. You could take your spacecraft, get out in space, and use a compass to find your direction! Even more strange, such a field would give everything that interacts with it a potential. If the field strength were the same everywhere, surely this would move all charged matter into one corner of the universe!

What is the problem? Basically, the electromagnetic field has a direction. When such a field is non-zero, space–time is no longer symmetric! There is a preferred direction and such a field is not Lorentz symmetric. We have never observed any violation of this fundamental principle, so this type of field cannot serve as a Higgs field. You may be confused about the fact that the electromagnetic field creates a direction in space–time. So why is that allowed? The electromagnetic field is always created by a source. And the direction of the field is always related to that source. So, in that case, the presence of the source breaks the symmetry of space–time. However, this differs from a field that is non-zero *everywhere*: in that case, *all of space–time* would be no longer symmetric and such a global violation of symmetry has not been observed.

What to do? We must find a different type of field to represent Higgs. From the prior discussion it is clear that it must be a field that does not have any direction. Such a field is called a scalar field. It can have a value, but no direction. By comparison, the electromagnetic field is a vector field: just like a vector, it has a value and a direction.

So, Higgs must be a scalar field. Other than that, it is not clear what constitution the Higgs field should have. One can think of a minimal constitution containing only one scalar field. But there is also a constitution containing four scalar fields (Lorentz invariance demands fields to be made of 1, 4, 16, ..., 4^n components). The latter seems to be favoured by theory, so let's check that one out.

In this theory, the four fields are called φ^1 , φ^2 , φ^3 , and φ^4 . As a scalar field, Higgs has spin 0 [Ref. 15]. Each field can be excited. Hence, we have four excitations to go with the four field components:

- A_0
- H^+
- H^-
- H_0 .

These Higgs fields have certain characteristics. First, a single scalar field cannot have a phase (just as one number cannot have a phase), but Higgs must have a phase in order to be able to have a charge: without phase, we cannot create a phase change. We can construct a superposition of two Higgs fields that together can have a phase. So, we need φ^1 and φ^2 to be in such a superposition and we need φ^3 and φ^4 to be in another such superposition. This will prove important later when we discuss how such a superposition behaves.

Each superposition can turn into the other. Consequently, the individual superpositions obey an SU(2) symmetry and carry isospin charges of -1 and $+1$. Both superpositions can produce phase shifts, so they carry a weak hypercharge of $+1$.

Now you may feel that all these characteristics of Higgs come falling out of the blue sky. And you are right. The Higgs field is assumed to be like this. The assumption is made so that it will have a property that allows a particular type of symmetry breaking. And it is deliberately connected to U(1) and SU(2) so that it can change these fields when its symmetry breaks. That is why we need the four Higgs fields, so that we can give it an Isospin charge and a weak hypercharge.

So, let's check out how all that works. To start with, there is a rather special mass potential connected to the Higgs field. In the early universe, Higgs was the only field that was massive. What caused it to be massive is unclear. What is more interesting is that the mass potential comes with an opposite sign. This means that it is a spring that works exactly the other way around compared to the mass springs we have seen so far. What does that mean? When we looked at the rope + springs model to explain what mass is, we considered springs that pull the rope. So, any field amplitude would have to pull the springs up. When the springs work oppositely, they rather push the rope (see Fig. 16.3). Hence, the field is non-zero. The only thing counteracting this push is the potential that comes from the elasticity of the rope. So, the elasticity of the vacuum works against the mass spring of the Higgs field.

In the usual situation, the field is at rest when the field strength is 0. After all, the springs that pull the rope are at rest and the elasticity of the rope would itself also be at rest. For Higgs this does not work that way (see Fig. 16.4). The springs pushing the Higgs field out of its rest position lead to a potential that is not 0 when the field strength is 0. In fact, to get a 0-field amplitude, energy must be put in, pushing the springs back to 0. This is the mass energy of the Higgs. The potential can become 0, but this we find at a spot where the field pushes against the springs and they are in equilibrium. Hence, the field is not 0 at that spot. In the picture we also see that there are now two places in the potential where the potential is at its lowest position. These two situations correspond to the two examples on the right of Fig. 16.3. One shows

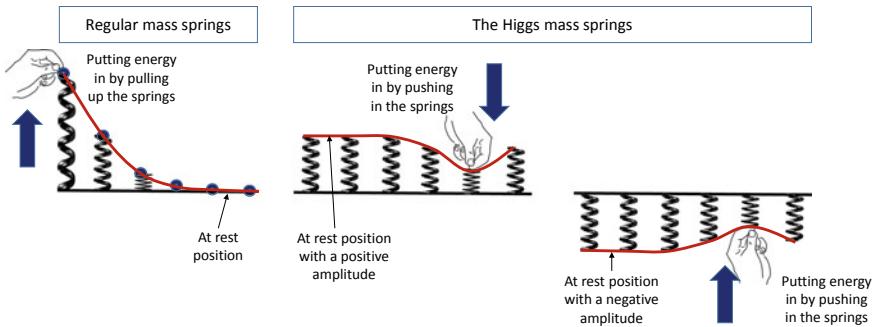


Fig. 16.3 The usual mass spring and rope (left) and a Higgs mass spring and rope (right, two examples). You can see that there are two “rest” positions on the right: one with negative field amplitude and one with positive field amplitude. In order to get a 0 amplitude, one has to put energy into the rope to push in the springs

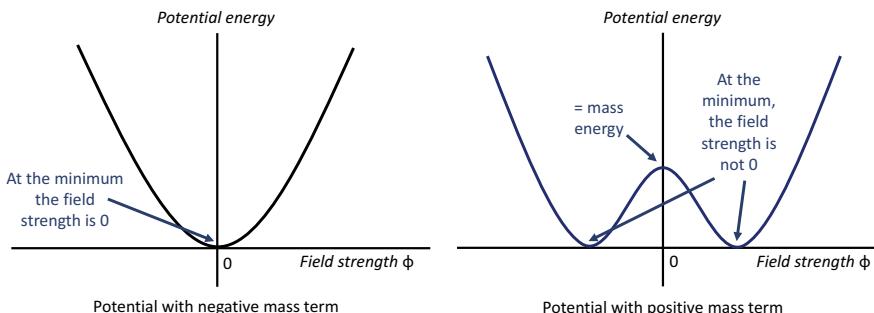


Fig. 16.4 A potential with a negative mass spring (left) and a potential with a positive mass spring (right). A negative mass spring is a spring that pulls the rope when it gets an amplitude. A positive mass spring pushes the rope away from its zero-amplitude position. As a result, the rope will have an amplitude in its rest position and needs to make some effort (put in some energy) to push in the springs in order to have a zero amplitude

an equilibrium with a positive field strength and the other shows an equilibrium with a negative field strength. This will prove essential in what follows.

The positive mass potential creates a “bump” in the field potential. The consequence is that there is not one potential minimum, but two. And both potential minima are found at positions where the field is not zero.

Higgs consists of two superpositions, each of which is made out of two field components. So, let’s consider one of its superpositions, and take a look at its potential. Since it consists of two field components, we must take the potential in Fig. 16.4 and apply it to both field components. This leads to the so-called Mexican hat potential (see Fig. 16.5).

The ground state forms a circle on which in most cases both field components are non-zero. Hence, the ground state (the lowest potential) requires at least one of the

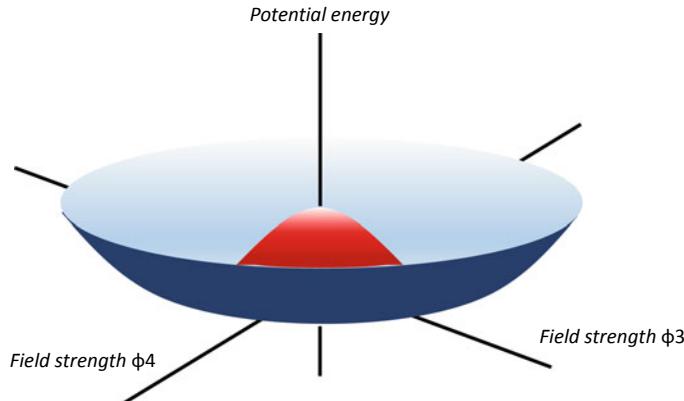


Fig. 16.5 Mexican hat potential for two Higgs fields in one of the superpositions

field components to be non-zero. So, when the Higgs potential is at its lowest, the field must be non-zero everywhere in space.

We are going to use all this information when we break the symmetry of Higgs, so hold on! Then it will become clear how the assumed Higgs field changes everything! First let's sum up what we have before the symmetry breaking.

16.3.1 Fields Overview First 10^{-12} s

When we extend the overview of fields with the Higgs field, we get Table 16.7.

Basically, only one row has been added, showing the Higgs. We did not enter all the Higgs fields, in order not to clutter the overview. Note also that Higgs does not yet interact with the fermions.

Table 16.7 Some fields and wave types in the early universe and their mutual interactions

Field	Wave type	Mass “spring”	Interaction “spring”
Right-handed electron	Fermion	0	With B
Left-handed electron	Fermion	0	With B, W1, W2, W3
Neutrino	Fermion	0	With B, W1, W2, W3
B-boson	Gauge boson	No	With charged fermions and bosons
W1	Gauge boson	0	With γ_e , e (left), W1, W2, W3
W2	Gauge boson	0	With γ_e , e (left), W1, W2, W3
W3	Gauge boson	0	With γ_e , e (left), W1, W2, W3
Higgs	Gauge boson	Opposite	With B, W1, W2, W3

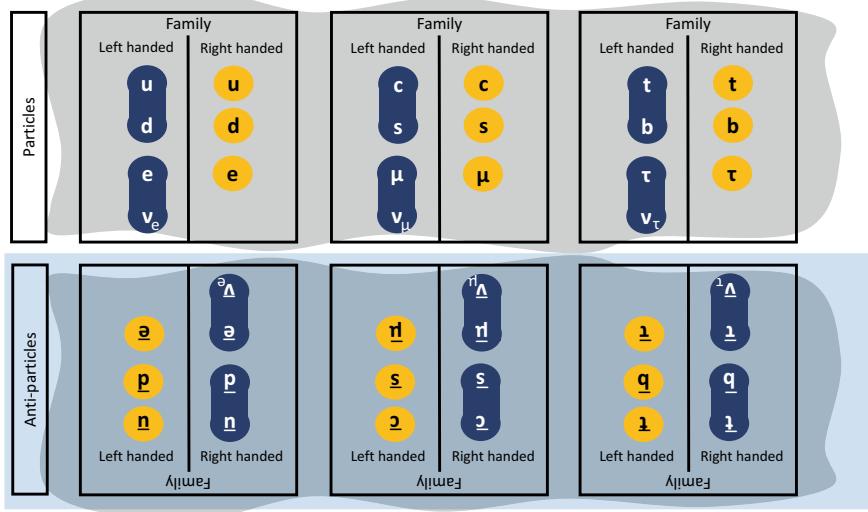


Fig. 16.6 Overview of fundamental particles with U(1) symmetry (grey) and SU(2) symmetry (bridge between the particles that are two faces of the same particle)

Let's see what the overview of fundamental particles looks like now in the early universe. After that we will take a look at how it all changes after the first 10^{-12} s.

16.4 Fundamental Particle Overview 2

In the overview shown in Fig. 16.6, we see that the isospin SU(2) symmetry between some particles (blue) actually made them two faces of the same particle (blue connector). Also visible is the clear distinction between left-handed particles and right-handed particles, as only left-handed particles are subject to the SU(2) symmetry. At this high energy stage, the Higgs field is zero everywhere and therefore does not do anything yet.

In the overview the grey area indicates which particles are subject to the U(1) symmetry. Before symmetry breaking, the U(1) symmetry applies to every fermion. Hence, the grey area is wider than in Fig. 15.2 and covers all particles. All fermions have a hypercharge. This must be so, since all particles should feel phase shifts, as required by Lorentz invariance. This invariance requires a phase shift to be carried away and there is no reason why one fermion should be able to produce phase shifts and not the other. We will see later that symmetry breaking will lead to a different situation. In that case Lorentz invariance will require a U(1) symmetry that does not apply to all particles.

For anti-particles we see that the isospin SU(2) symmetry applies to right-handed chirality. This can be understood as follows. When we turn particles into anti-particles, it is not only time that gets reversed, but also parity (we will get back to this in Sect. 17.5 when we discuss CPT symmetry). And a parity reversal also swaps the chirality (see Sect. 13.4). So, when particles experience an SU(2) symmetry on their left-handed versions only, this must be so for the right-handed versions of their anti-particles.

Chapter 17

Symmetry Breaking and the World Was Never the Same Again



Symmetry breaking is no stranger in the world of physics. Let's look at some examples closer to home in order to get a feel for what symmetry breaking entails.

Let's look at the atmosphere once more. Hot air can contain more water than cold air. Consequently, in a hot summer week the level of moisture can build up in the hot air. The moisture is equally spread throughout the air, hence symmetric in all directions/locations. When cold air arrives, the air must cool down. The air cannot contain all the moisture and some of it must condensate. Hence, droplets start to form. As of that moment the symmetry is broken: there is a difference between droplets and air. Droplets get together to form clouds and eventually rain. Cloudy air is clearly no longer symmetric.

The same goes for a plasma such as in the sun. The temperature in the sun is high enough to strip atoms from their electrons. Hence, a plasma is formed that consists of loose electrons and protons (assuming for the moment that there is only hydrogen in the sun, which is not true, but also not important for the argument). This is a symmetric situation, like the hot air. When that plasma is cooled (e.g., when it is ejected from the sun in a solar flare), the electrons recombine with the protons to form hydrogen. The symmetry gets broken.

One final example. We discussed the orbits of electrons in the atom in Sect. 13.1. Orbita with the same orbital momentum have the same energy. These energy levels are said to be degenerate, since they cannot be distinguished by their energy. Hence, there is an orbital symmetry in the atom. However, we also saw that the cloud of virtual photons and electron–positron pairs of dressed electrons interacts with the protons in the nucleus. The result is that the energy levels of degenerate orbits start to differ: they are no longer degenerate and the symmetry is broken.

So, in general we see in these examples that symmetry gets broken as a consequence of interactions (e.g., the atom) or as a consequence of lowering the energy level of the system (e.g., air, plasma).

17.1 Mixing Fields

When we explained what a field was, we also used the moisture and air example. So, the atmosphere has different fields such as the moisture field. The field values are equal to the moisture level. Let's introduce another "field": that of aerosols (or dust particles). The atmosphere is full of those and they play an important role in forming raindrops. So we now have two types of field in the atmosphere and they live independently. They are not mixed in the sense that they can be considered separate fields that do not influence each other and do not look like each other. They exist separately throughout the carrier of the fields: the atmosphere.

Now suppose the temperature drops below the point where the moisture level cannot be contained in the air. When the aerosol field is zero (i.e., there are no aerosols present), nothing happens! The air gets "super cooled". This is a state where the moisture would like to condensate, but it cannot. Moisture requires a surface to condensate on before it can form droplets. When the aerosol field is not zero, droplets can form on the surface of the dust particles. What happens then is that a new field gets born! We call that field the "cloud field". It is the field of water droplets floating in the air. This field cannot exist without the two fields it is born from and it cannot exist above the symmetry-breaking temperature (see Fig. 17.1).

The interesting thing is that the new field is a mixture of the old fields, moisture and aerosol, but it has new characteristics. These characteristics are partly inherited from the old fields. For instance, the field contains dust as well as moisture. It is just that they exist in a combined form that displays different properties. For example, an electromagnetic field (e.g., light) can penetrate a moisture field and an aerosol field without being scattered much. It does not interact with the moisture field, and only

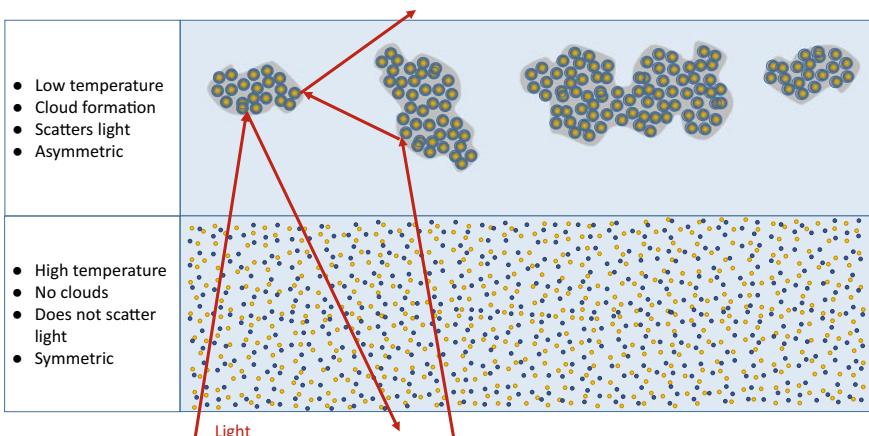


Fig. 17.1 Symmetry breaking. The lower part shows a symmetric mix of dust (yellow dots) and moisture (blue dots) and the temperature is high. The upper half shows a set of blue balls with a yellow core, asymmetrically grouped in clouds. The temperature is low. Light is reflected from the clouds, but it is not in the lower part

slightly with the aerosol field. But the cloud field is different: light gets scattered in a cloud field. So, it interacts heavily with it. Something similar goes for a sound field. This penetrates a moisture and aerosol field at a velocity of approximately 343 m/s. This velocity depends on the air density and temperature. In a cloud field, the velocity of sound is slightly higher because the density of the air is higher. So, excitations in both the electromagnetic field and the sound field (photons resp. sound waves) behave differently in a cloud field. This is a consequence of the fact that they “interact” differently with the cloud field compared to the moisture and aerosol fields.

This example shows that fields can mix and form different fields with different properties. It also shows that symmetry breaking can cause fields to mix. When the temperature is high enough for the air to contain the moisture (before symmetry breaking), the cloud field has a zero value everywhere. After symmetry breaking, it has a non-zero value (but not necessarily everywhere). Suppose you take off in an airplane on a cloudy day. At first you can see the ground from the window. The temperature of the atmosphere is high enough to contain all the moisture. But when you go higher, the temperature drops. At some point the plane enters the cloud field, and you can no longer see the ground: symmetry breaking has taken place!

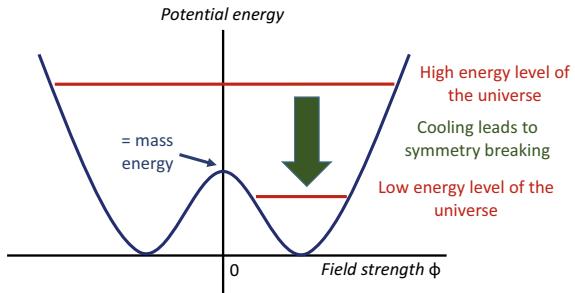
Why is this important? We will see in this section that the Higgs field can be compared to the moisture field. Below a certain temperature, its symmetry breaks. When that happens, Higgs mixes with some fields and leaves a non-zero field that interacts differently with other fields. One of the consequences is that these other fields will gain mass from that interaction. Just as the degenerate energy levels of the electron orbits in an atom get differentiated in energy as a consequence of interactions, the interaction with Higgs does that for the massless and degenerate W-bosons.

Keep in mind that this is a *very rough comparison*. There are *many differences*. The fields we find in the vacuum are in no way similar to the fields we described in the atmosphere. However, the example does show us that fields can mix and become different fields as a consequence of symmetry breaking. This idea might make it easier to accept the field mixtures we will discuss in the following sections.

17.1.1 What Condensates in the Vacuum?

So how does that work for the vacuum? The vacuum has a temperature too. What makes up that temperature? The current temperature of the vacuum is about 2.7 K. The heat that goes with that temperature is stored in the waves that are present in the vacuum. Electromagnetic radiation is mostly responsible for those waves. This type of radiation is called “black body radiation”. A black body absorbs radiation perfectly and does not reflect any. So, the radiation emitted by a black body originates purely from its temperature. Temperature is nothing else than moving and shaking atoms in the black body. Such movements of charged particles produce radiation. When a black body is in equilibrium with its environment, it absorbs as much radiation as it emits. Hence, it heats up as much as it cools down and the net effect is that

Fig. 17.2 Symmetry breaking in the early universe



it maintains the same temperature. Suppose the black body is put in the vacuum of the universe. The temperature of such a black body in equilibrium with the vacuum is equal to the temperature of the vacuum. In the current universe, the black body would have a temperature of 2.7 K (far away from any stars), which we consider to be the temperature of the vacuum.

In the past the temperature of the vacuum was higher. We saw before that the universe would break symmetry at around 10^{16} K. To understand what happens we need to look at the Higgs potential (see Figs. 16.4 and 16.5). Before symmetry breaking, the temperature of the vacuum was so high that the Higgs field, when in equilibrium with the vacuum, had an energy level above the bump in the potential (see Fig. 17.2).

Consequently, the Higgs field was excited to an energy level so high that it could be symmetric! It was in a state comparable to “hot air containing a lot of moisture” or to a plasma. When the temperature dropped below 10^{16} K, the energy level of the Higgs field had to drop below the bump. So, what happened there? Basically, the field had to make a choice whether to be excited in the left potential well or in the right potential well. It could not do both. The field had to make a choice, but how did it do that? And what were the consequences of this choice? This will be the subject of the present section.

As soon as the field has made its choice, the symmetry is broken, since the field values will be positive (in Fig. 17.2) if it chose to go to the right or negative if it chose to go the other way. In the ground state (at very low energy) the field will not be 0. It has some non-zero value.

Another important issue is that the height of the bump signifies the mass of a Higgs excitation. When the field is 0, the positive mass potential is the only “spring” that remains. Hence, the energy level when the field is 0 equals the mass of the Higgs excitations. When the field has an energy above that level, it contains Higgs excitations just as moist contains water in hot air. When the temperature cools below that energy level, the Higgs field cannot contain Higgs excitations any longer and they will need to “condense” into Higgs particles: excitations that have a higher energy than the average field. The total energy of the remaining field + excitations is of course in equilibrium with the vacuum.

So, Higgs is the only field that breaks its symmetry when the temperature of the vacuum drops. Still, this will have major implications for the other fields as well. This is caused by the fact that Higgs interacts with these fields and, upon symmetry breaking, that interaction changes as well.

So, let's dig a little deeper and see what surprising processes that will bring us.

17.2 Breaking the Symmetry of the Higgs Field

We break the symmetry by lowering the energy level. We have four field values to consider. Let's first take a look at the Mexican hat potential for two of those field components in their superposition. This is a more complicated situation than the one-dimensional case. We can now have particles (field excitations) move in complicated ways through the potential (see Fig. 17.3).

However, such a movement can be dissected into two basic movements: one in the “gutter” of the hat and one sideways. Viewed from above, this looks like Fig. 17.4. The excitation that moves through the gutter does not have to climb the potential at any point, so it does not have energy (and hence does not have mass). Looking at this as a separate type of excitation, it is a particle that is called a “Goldstone mode” or “Goldstone boson” [e.g., Ref. 9, p. 228; Ref. 8, p. 240]. The other type of motion goes up and down the potential, just like a harmonic oscillator. This excitation does have energy and is a more “normal” type of excitation. So, we see that a regular Higgs excitation would break into two different types of excitation when the energy of the field drops below the symmetry breaking point: one (massless) goldstone boson and one regular (massive) boson. The mass of that boson is equal to the energy of the first excitation level.

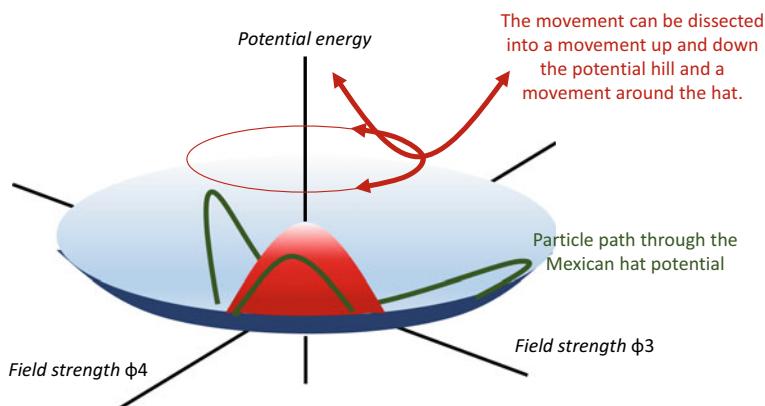


Fig. 17.3 Particle moving in a complicated way through the Mexican hat (green), and the way this movement can be dissected into two basic types of motion (red): up and down the potential hill and around the hat

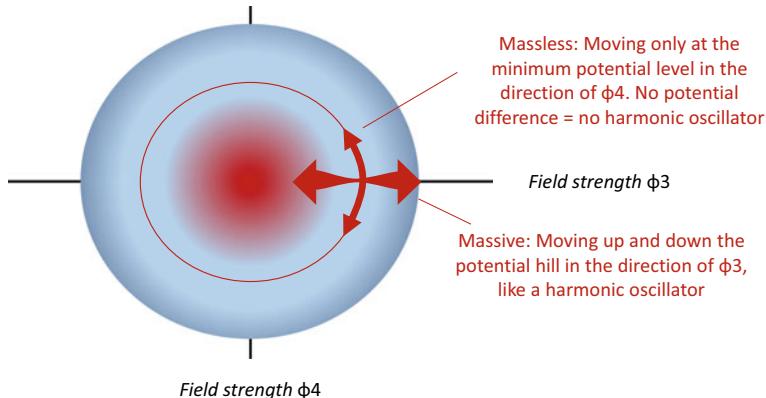


Fig. 17.4 The basic movements seen from above

Now, you may wonder why that makes this boson massive, since this is not the case for, e.g., photons. Photons do have excitation energy, but they don't have rest mass. Their energy levels originate from the elasticity of the vacuum alone. Fields gain mass when extra springs are added. The Higgs field was already massive before symmetry breaking, indicating that it has its own "extra" springs baked into the Higgs potential. So, when a Higgs boson goes up and down that potential it will feel those springs and become massive.

From Fig. 17.4 we see that one field can have a 0-field value at the lowest energy point. In that case the other field has a value that determines the total field strength. We can extend the idea for two field components to four field components. For the four field components of the Higgs field, we can make a similar choice where only one of the field components has a non-zero value and the other three are 0. The three "zero" fields cannot get excited since their "excitations" will move in the gutter of the potential. So, these "excitations" (goldstone bosons) cannot become harmonic oscillators. Consequently, they will not hold or transport energy, which means that they do not have an impact anywhere. If they did interact with any field, that interaction would have zero impact.

So far, we cannot identify the direction in which the symmetry gets broken in the Mexican hat. Nature could potentially choose any direction. So why would we choose a particular direction in which three of the fields are 0? Looking at the consequences gives us an answer to this question. Again, this is determined by what leads to the universe as we observe today.

17.2.1 Consequences for the $U(1)$ and $SU(2)$ Gauge Bosons

Our starting point is that we take a broken symmetry in which three Higgs fields are zero and one is non-zero. Consequently, there can only be a wave in one of the Higgs

fields: the non-zero field. This wave is an excitation of that field and we can associate it with the Higgs boson. This is a massive excitation. Furthermore, the non-zero field is all pervasive, existing throughout the vacuum.

The Higgs field interacts with all the bosons W1, W2, W3, and B, but not in the same way. Since three of the fields are 0, the interaction with those fields does not lead to a potential for the bosons. However, the field that is non-zero does provide a potential in the interaction. Remember that the interaction potential is determined by the coupling constant (which is called g in this case) and the field strengths. So, when the strength of three of the Higgs fields is 0 and there are no excitations in these fields, they will not lead to a potential or interaction.

The interaction with W1 and W2 is straightforward. It is not an interaction with an excitation in the Higgs field. It is a straight interaction with the non-zero field itself. This is different from the interactions with other fields we discussed before: such an interaction can only be an interaction with an excitation in those fields. That means that only when there is a wave in those fields can an interaction with such a wave be felt. But when there is no wave, the field is 0 and no potential is felt. In the case of Higgs, the field is non-zero. So, an interaction with that field is a continuous interaction, and not limited to when there is a wave in the field. This interaction is like moving through a continuous medium, like our example of light moving through glass. So here we see the continuous Higgs potential appear, providing the homogeneous springs throughout the universe that give mass to W1 and W2. The actual value of the mass of W1 and W2 depends on the coupling strength g , which is the same for both particles.

Things are different for W3 and the B-boson. It turns out that they interact with Higgs only in a mixed state, rather as the moisture field could only condensate in a mixture with the aerosol field. So, Higgs connects to a mix of W3 and the B-boson. W3 and B have the same hypercharge ($Y = 0$) and isospin ($I_z = 0$). In a way, they are like the two degenerate states of one particle. So, we cannot distinguish between them in the mix. Consequently, there are two possible mixes: a symmetric mix and an asymmetric mix. The symmetric mix turns out to be massless, while the asymmetric mix is massive. So, Higgs interacts with W3 and B in such a way that it mixes them up in a massless mix and a massive mix. The massless mix operates as the photon in our universe and the massive mix as the third weak boson. That boson has been renamed as the Z° -boson.

The coupling constant for the photon and Z° are determined by the mix. Let's say that the mix is set up in the following way:

$$Z^\circ = g W3 - g' B \text{ (asymmetric)}$$

$$\text{Photon} = g' W3 + g B \text{ (symmetric)}$$

The photon mass must be 0 (since we measure it to have 0 mass in our universe). However, Higgs interacts with W3 and with B. So how can the photon mix be massless? Moreover, Z° could not be massive if Higgs did not interact with at least

one of them. So, the only way for the photon to be massless is when the mix itself is 0:

$$g' W3 + g B = 0$$

This can tell us what the mixing ratio between $W3$ and the B -boson is for the photon! For the photon term to be 0, this mixing ratio must be:

$$\text{Mixing ratio } W3/B\text{-boson} = -g'/g$$

So, the mixing ratio depends on the coupling strengths to the Higgs field when we demand that the photon mass equals 0. As a consequence, we can also determine the mixing ratio for the Z° .

So, we see that the old $SU(2)$ $W3$ -boson gets mixed with the old $U(1)$ B -boson. New particles emerge from this mix. The mixing away of the B -boson has an implication for the charge it represents. That charge has changed and in fact mixes as well. The new charge is that related to the photon, but the photon is a mix of $W3$ and the B -boson. Hence, the new charge depends on the mix and consequently it depends on the mixing ratio g'/g .

So, what determines the mixing ratio? You can imagine that $W3$ and B couple differently to the 4 components of the Higgs field. But the values of the components of the Higgs field depend on the direction of the symmetry breaking in the Mexican hat. Hence, the coupling strengths of $W3$ and B to the Higgs field depend on that choice of direction as well. Consequently, different choices of direction lead to different values for g and g' and hence a different mixing ratio. There is one requirement for this process: the photon combination of $W3$ and B must treat the vacuum as symmetric. The reason is that it is a $U(1)$ symmetry that is *Lorentz invariant in the current universe*. So, nature's choice of direction has led to precisely this feature. Now we can ask: what choice of direction makes the present-day $U(1)$ symmetry Lorentz invariant? There is only one direction that gives us this feature: the one where three fields are 0 and one Higgs field is not! This choice mixes the $W3$ and B into a Lorentz invariant photon in just the right way.

This choice of direction leads to a combination of $W3$ and the B -boson that treats the vacuum as symmetric. This combination (the photon) also has the feature that it does not interact with Higgs! The mixing angle that goes with this choice of direction links the charge of the photon to the old hypercharge Y and old isospin I_z :

$$Q = I_z + Y/2$$

- The vacuum turns out still to be invariant with respect to a $U(1)$ transformation over an angle equal to $Y/2 + I_z$, as if the new charge Q has become the $U(1)$ transformation.
- You could find this formula yourself by trying out all particles with their I_z , Y , and present-day Q . Only one formula fits. The right-handed electron has a hypercharge

of -2 , but today it has a charge -1 . So, the charge should be $Y/2$. The left-handed electron had $Y = -1$ and $I_z = -\frac{1}{2}$ and its present-day charge must be -1 as well. So $I_z + Y/2$ would work. This works for all particles.

Since the photon combination treats the vacuum as symmetric, the other combination (Z°) consequently cannot. The reason for this is that the Z° combination is a different combination. There can only be one combination that treats the vacuum as symmetric. Clearly, the combination that treats the vacuum as not being symmetric fits a carrier of the weak force, as the weak force does not treat the vacuum as symmetric either. The weak force treats particles with left and right chirality differently. It takes a parity operation to change left-handed chirality into right-handed and vice versa (see Sect. 13.4). The parity operation exchanges all directions with their opposite directions. When a force distinguishes between a chirality and its parity-exchanged version it clearly does not treat space as symmetric.

Since the photon treats the vacuum as symmetric, it must also treat left-handed and right-handed fermions as equal. Therefore, the electromagnetic charge of all left-handed and right-handed fermions must be equal. However, they do not have an equal hypercharge! So, before symmetry breaking $U(1)$ did not treat the vacuum as symmetric!

After symmetry breaking, left- and right-handed fermions must have an equal charge. This shows again why the mix that leads to the photon must lead to the relation $Q = I_z + Y/2$. This relation makes the charges of left- and right-handed fermions equal. A consequence is that the neutrino is left with 0 charge! There is some logic in this since there is no right-handed neutrino. So, if the neutrino did have a charge, how could the right-handed neutrino have the same charge when it does not exist? Consequently, if the neutrino had a non-zero charge, there would be an asymmetry in $U(1)$ after symmetry breaking.

The three goldstone modes of the broken Higgs field no longer play a role. This is a consequence of them mixing with the formerly massless W_1 , W_2 , and the Z° mix of W_3 and B . So, we see that they disappeared while the interaction of the Higgs field with the W_1 , W_2 , and the Z° mix of W_3 and B changed and mixed these particles into new massive ones. In a sense, the way the W 's and B mixed with the zero fields of Higgs made them interact with the non-zero field, and this is why they became massive. The symmetry breaking of Higgs was such that no particles mixed with the fourth Higgs field component. Hence, the fourth Higgs field component remains. The choice of direction in the Mexican hat has made that field non-zero throughout the universe and its excitations are massive Higgs bosons. This is the Higgs excitation that is associated with the original H° field. The degeneracy of W_3 and B is now removed by mixing them into Z° and the photon.

The Higgs boson has been found by CERN after a long search. It took such a long time because the particle is very massive. CERN's site states "On 4 July 2012, the ATLAS and CMS experiments at CERN's Large Hadron Collider announced they had each observed a new particle in the mass region around 126 GeV. This particle is consistent with the Higgs boson predicted by the Standard Model." [Ref. 15]. As it is

Table 17.1 The mixing of fields as a consequence of symmetry breaking in the Higgs field

Old boson	New particle mix	Mix with Higgs field	Result after symmetry breaking
W1		H^+ and W1	Massive W^+
W2		H^- and W2	Massive W^-
W3	$Z^\circ = W3 - B$	$A0$ and $(W3 - B)$	Massive Z°
B	$\text{Photon} = W3 + B$		Massless photon

a scalar particle, the Higgs boson is the first fundamental scalar particle discovered in nature.

The mixing process is called the Higgs mechanism. We summarize the mixing process in Table 17.1.

The non-zero Higgs field also interacts with fermions. In the same way it gives them (bare) mass and the size of the mass is determined by the size of the coupling constant. The fact that the Higgs field is non-zero everywhere is responsible for that: it leads to a continuous and homogeneous interaction.

17.2.2 *Mass of the W^- , W^+ and Z°*

Henceforth, we will use the post-symmetry-breaking names for W1 and W2: W^+ and W^- . From experiment we know the masses of $W^{+/-}$ and Z° to be:

$$\text{Mass } Z^\circ = 91.19 \text{ GeV}$$

$$\text{Mass } W^{+/-} = 80.40 \text{ GeV}$$

These masses are very high. The Z° is for instance approximately 97 times as heavy as the proton. Let's compare this to gold. A gold atom consists of 197 Nucleons (protons and neutrons). It weighs about 183 GeV, and so the Z° weighs as much as half a gold atom.

We have seen that the weak force originates from turning one face of, e.g., an electron into the other face, a neutrino. When the electron makes such a change, it produces a W. After symmetry breaking, this process changes. We will see later what types of interaction this leads us to. With regard to the mass, just try to imagine that an electron (weighing only 0.5 MeV) has to produce a virtual W of 80 GeV when it turns into a neutrino. For this process to happen, the electron needs to rely entirely on the uncertainty relations. The W cannot be produced out of its own mass or energy.

The electron produces a virtual photon cloud around it. Taken from the mass of the W's, it will be much less enthusiastic about producing a virtual W-“cloud” as well.

However, in principle it works the same way. The electron can turn into a neutrino, producing a W, for the very short time an extra energy can be “borrowed” according to the uncertainty relation:

$$\Delta E \Delta t \geq h/4\pi$$

Since the amount of energy needed to produce a W is so high, we are talking about a major loan here. That can only exist during a very short time. This time is of the order of 10^{-24} s. At a maximum velocity of c (3×10^8 m/s), this means that the weak force has a range of about 10^{-16} – 10^{-17} m. At such a distance, the weak force is about 10,000 times weaker than the electromagnetic force. But at 10^{-18} m it is already as strong as the electromagnetic force. Consequently, we do not notice this force in daily life. It only plays a role below the scale of a nucleus. The size of a nucleus is typically about 10^{-14} – 10^{-15} m.

A different way of looking at this is to consider the W that is produced as a disturbance rather than a proper wave. A low energy disturbance must be very far from the resonance point (where a real W could be created), which makes it very unlikely to be produced and hence very short-lived. One way or the other, we arrive at the same result: the weak force is weak because of the huge mass of the W’s and Z.

When an electron turns into a neutrino, producing a W-disturbance, it can reabsorb the W and turn back into an electron. Alternatively, the W-disturbance can be picked up by another particle making such a change and the electron will come out as a neutrino.

17.2.3 Consequences for the Fermion Interaction Potentials

We have not yet discussed the interactions of the fermions with the U(1) and SU(2) gauge fields. Before symmetry breaking, we still have the interaction potentials of the electrons and neutrino with the U(1) B-boson (hypercharge) and the SU(2) W1, W2, and W3 bosons (isospin).

As a next step we use the equations showing how these particles get mixed in order to replace W1, W2, W3, and B in the interaction potentials. This means eliminating the old boson fields and replacing them with the new fields, but also shows how fermions couple to the new fields $W^{+/-}$, Z° , and A (photons). In particular, the mixing of W3 and B into the photon and the Z° causes many changes. This results in the electric charge being a mixture of hypercharge and isospin, as we have seen. But it also results in the coupling to Z° being different from the coupling to $W^{+/-}$. The latter two couple to isospin as before, but Z° couples according to a mixture between charge and isospin. This mixture is determined by the mixing ratio, just as we saw before with the photon. We will not discuss the resulting Z° charge as we will not be using this further. The result is summarized in Tables 17.2 and 17.3.

Table 17.2 The new charges of the fermion fields as a consequence of the symmetry breaking of the Higgs field

Left-handed chirality	Y_w	I_z	$Q = I_z + Y/2$	Right-handed chirality	Y_w	I_z	$Q = I_z + Y/2$
$\gamma_e, \gamma_\mu, \gamma_\tau$ neutrinos	-1	+½	0	Neutrinos (probably) non-existent	x	0	x
e, μ , τ leptons	-1	-½	-1	e, μ , τ leptons	-2	0	-1
u, c, t quarks	+1/3	+½	+2/3	u, c, t quarks	+4/3	0	+2/3
d, s, b quarks	+1/3	-½	-1/3	d, s, b quarks	-2/3	0	-1/3

Table 17.3 The new charges of the boson fields as a consequence of the symmetry breaking of the Higgs field

Boson	$Q = I_z + Y/2$	I_z	Y_w	Mix of
Photon	0	0	0	$W3 + B$
W^+	+1	+1	0	$W1 + H^+$
W^-	-1	-1	0	$W2 + H^-$
Z°	0	0	0	$(W3 - B) + A^\circ$
Higgs	0	-½	1	H°

In Table 17.2 we see that the photon does not couple to the neutrino, unlike the B-boson (hypercharge), which did couple to the neutrino before symmetry breaking.

Another important result is that Z° couples to both the electron and the neutrino, but with different strengths. Since the Z-boson does not have charge or isospin, it can only transfer momentum, energy, and spin. Unlike the W^+ and W^- , Z° cannot change the charge of a fermion. This is why it is called the *neutral* current. “Current” here has nothing to do with electrical current, but refers to “isospin current”.

The neutrino can interact with other matter only through the weak force. When it does so through W^+/W^- , it is involved in changing the flavour of fermions. Only Z° offers the possibility for a neutrino to scatter off other fermions without turning them into another flavour. Z° allows for momentum transfer between a neutrino and a fermion. We call that process “elastic scattering”, since it does not change the particles involved. This is like a rubber ball that bounces off the floor. The ball and the floor do not change in this scattering process. “Inelastic scattering” would be more like bouncing a ripe tomato off the floor. Clearly that changes the floor (it gets dirty) and the tomato (squashed). Hence, Z° acts like a force, as we know from the electromagnetic force! The only (major) difference is that Z° is so heavy that the force has a very short range.

Neutrinos have the same effective Z charge as isospin. That means that they are as likely to scatter elastically through Z° (leaving particles the same) as to scatter inelastically through $W^{+/-}$ (turning particles into a different flavour).

Right-handed fermions (except neutrinos) have a non-zero effective Z^0 charge. This means that they can feel the Z -boson, while they cannot feel the $W^{+/-}$ bosons, since they cannot turn into neutrinos. This is a consequence of the Z^0 being a combination of the W^3 and B . So, anything right-handed having a hypercharge will have some interaction strength with the Z^0 .

17.3 Interactions

We have already seen some Feynman diagrams (see Chap. 10). The vertex is the point where a phase shift or an isospin rotation happens (in the case of electromagnetic or weak interactions). The phase shift or isospin rotation changes the particle that makes the shift or rotation, either in direction and momentum or by changing also its face. When it does so, it produces a gauge wave. That gauge wave is intended to propagate away the change in the particle wave. Consequently, at the vertex, the total charge and isospin on the left side of the vertex (the “before” situation) must be the same as on the right side of the vertex (the “after” situation). We have seen this before with respect to the momentum (“before” momentum = “after” momentum). In this section we will focus in the diagrams on the conservation of charge and isospin.

The photon has a limited number of ways to interact with matter and the interaction is simple. It leaves particles unchanged, except for momentum and (sometimes) creation/annihilation. The weak force is very different. A lot of different types of interaction are possible, often changing the face of the particles involved. All interactions are mediated by bosons, because bosons carry the phase shift or rotation that changes the fermion’s wave. There are no direct fermion–fermion interactions since fermions do not change another fermion’s wave directly. Hence, we can group the types of interaction around the type of boson. So, let’s dive into the different types of interaction that are possible and see what they involve. We begin with the photon.

17.3.1 Photon Interactions

A. Force-carrying interactions

A force is produced by a potential that is propagated from one particle to the other, bending the paths of both particles. Put more directly, momentum is transferred by the mediating boson. The photon mediates the electromagnetic interaction (see Fig. 17.5). This mediation does not change the face of the charged particles.

B. Creation /annihilation

The energy of a photon can be used to create a particle/anti-particle pair (pair production, see Fig. 17.6). This can also be viewed as a particle turning into its anti-particle

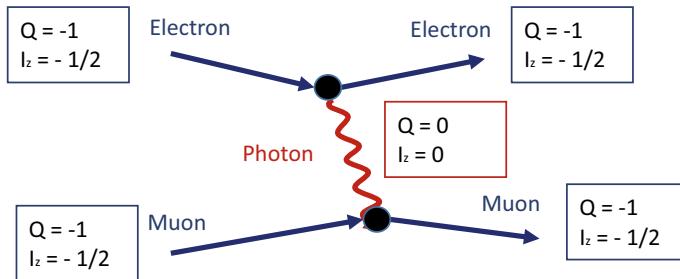


Fig. 17.5 Photon mediating the electromagnetic force

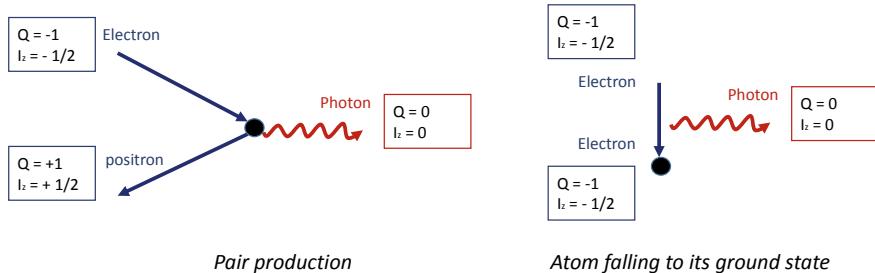


Fig. 17.6 The role of the photon in creation/annihilation processes

while producing or absorbing a photon. Other types of creation and annihilation are, e.g., an atom changing its energy level by creating or absorbing a photon.

C. Vacuum polarization

As discussed before, the photon can polarize the vacuum by splitting into a virtual electron–positron pair and reabsorbing it (see Fig. 17.7). During its existence, the electron–positron disturbance can polarize in an electromagnetic field.

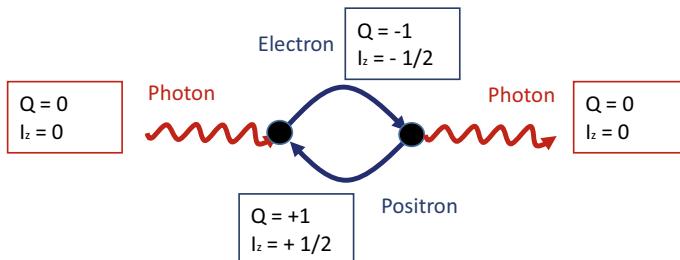


Fig. 17.7 Vacuum polarization

17.3.2 W-Interactions

A. Face-changing interactions

A neutrino can change its face into an electron, producing a W^+ disturbance (see Fig. 17.8). Charge can be added up at the vertex. The charge before was 0 (neutrino) and the charge after the rotation is -1 (electron) $+1$ (W^+). Hence, charge is conserved at the vertex. The same goes for isospin. The W^+ then gets absorbed by a down quark that changes face to an up quark. Here too, charge and isospin are conserved. Both quarks have to be left-handed, as the right-handed versions have 0 isospin and they cannot absorb the W^+ .

In the second diagram a down quark (L) turns into an up quark while producing a virtual W^- . The W -disturbance is absorbed by a positron that turns into a neutrino. Keep in mind that anti-fermions must be right-handed to interact with the weak force, as left-handed ones cannot.

B. Creation /annihilation interactions

Creation and annihilation processes involving W -bosons include face changing. For example, a top quark can annihilate against an anti-bottom quark to produce a W^+ boson (see the first diagram of Fig. 17.9). Now it becomes especially tempting to view a pair annihilation as a particle turning back in time, i.e., turning into its anti-particle. In the case of W -bosons they not only turn back in time, but also turn into their other face when they produce a W -boson. Note that the total charge on the left is $+1$ and, on the right as well. If we think about a top quark going in and producing

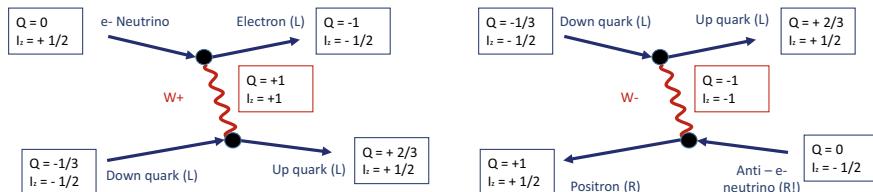


Fig. 17.8 Face-changing interactions involving W -bosons

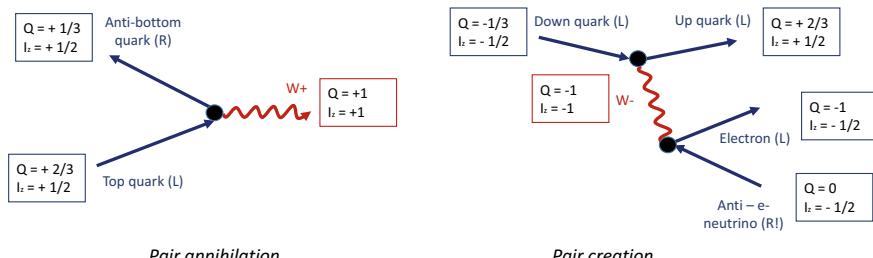


Fig. 17.9 Pair annihilation and creation interactions

a W^+ , the top quark must change into something with charge $-1/3$ (bottom). Just don't forget to change the sign of that charge when the end product is an anti-particle (anti-bottom)!

The second diagram shows the same but now with a W^- boson ceasing to exist and creating an electron and anti-neutrino pair. Here too, this could be viewed as the anti-neutrino absorbing the W^- and turning into its other face, as well as going in the other direction in time: the electron. Note also that this diagram is the same as the second diagram in Fig. 17.8, except that we have an electron leaving instead of a positron entering! The two processes are essentially the same, but viewed from a different frame of reference (different relative velocity).

The same could go for other diagrams such as the first diagram in Fig. 17.9. Instead of an anti-bottom quark entering, we could have a bottom quark leaving the diagram and it would just be a face-changing interaction. It all depends on the relative velocity of the observer.

C. Decay interactions

A muon cannot turn into another (non-muon) type of neutrino and it cannot just split up into an electron and something else. Still, a process is observed in which muons decay into electrons and two neutrinos. How does that work? First, a muon can turn into its muon-neutrino face and produce a W -disturbance (see Fig. 17.10). The W^- then decays into an electron and an anti-neutrino of the electron type.

The same goes for a pion (π^+), which is a composite particle made from an up quark and an anti-down quark. The up quark turns into the anti-down quark while producing a W^+ disturbance. The W^+ boson decays into an anti-muon and a muon type neutrino. Or put another way, the anti-muon absorbs the W^+ and turns into its neutrino.

The funny thing is that a pion (that is, π^+) is apparently made from a quark and the anti-version of its other face. Put differently, the pion consists of a quark pair that can turn into each other under the weak force. The plot thickens if we think about observing a pion from a different frame of reference, where the anti-down quark is just a down quark and the W^+ is produced by a rotation from up to down. Can we then consider the pion to be a particle? To answer this, we need to include the colour

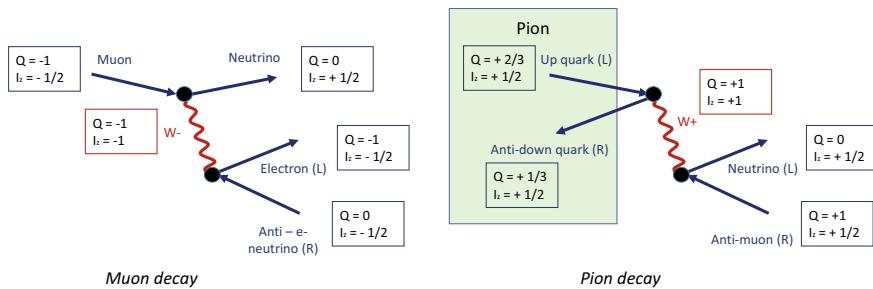


Fig. 17.10 Decay processes

charge of the quarks and we will get to that in the next chapter. For now, we can say that a loose quark cannot exist, so for a quark to change face it must be part of a composite particle that contains another quark or two. Consequently, the pion can consist of an up quark + anti-down quark that can co-exist at the same time (and are connected by the colour force), but not in the form of a loose up quark that changes into a loose down quark. Taking that further, the velocity of the up quark and the anti-down quark must be the same. Then, there is no reference frame in which the up quark remains an up quark, but the anti-down appears as a down quark. If you find yourself in a reference frame where the anti-down quark is actually a down quark, then the up quark must appear as an anti-up quark and you have a π^- (see the table for the particle zoo)! The π^- happens to be the anti-particle of the π^+ . And so, everything is consistent again!

D. Force-carrying interactions

The interactions discussed under A–C also have an impact on momentum, but they do not leave the particles the same. Hence, the force aspect of these interactions gets overshadowed by the face-changing aspects. There are also interactions involving W-bosons that produce the same particles as the ones that entered the interaction. In such processes the force aspect comes more to the fore.

Take for instance an electron that turns into a neutrino while producing a W^- disturbance (see Fig. 17.11). The virtual W^- can be absorbed by a neutrino that turns into an electron. Then the same particle types enter the interaction as the ones leaving the interaction. The momenta of the electron and neutrino can be differently distributed between them before and after the interaction. Effectively, the interaction leads to a momentum transfer. Alternatively, we can say that the electron scatters off a neutrino. The force that is created by this type of interaction is called a charged current, since it is carried by a (isospin) charged W-boson. It is an inelastic type of scattering as the particles change in the process.

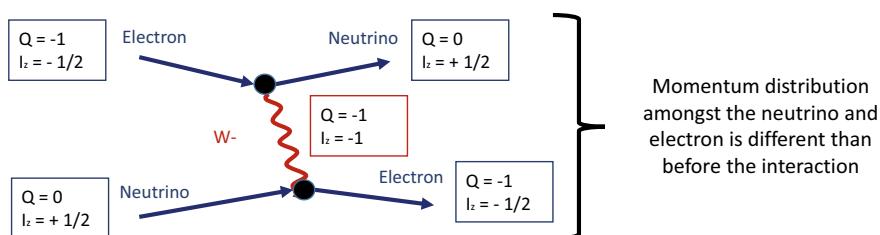


Fig. 17.11 Effective force created by a charged current process

17.3.3 Z-Interactions

A. Force-carrying interactions

The Z-boson can also be used to create a force. The Z-boson does not change the face of a fermion. Put differently, when a fermion turns into itself, it produces a Z-boson. A fermion that absorbs a Z turns into itself. So, we can look at the following process (see Fig. 17.12): a neutrino turns into itself, producing a Z disturbance. The Z disturbance is absorbed by an electron that turns into itself. All the Z has done is transfer momentum and energy. In this sense it looks a lot like the photon. Since no particles change in the process and Z does not carry charge, this process is called the neutral current. We can say that an electron scatters off a neutrino this way. In this case it is an elastic type of scattering since the particles do not change.

It is interesting to measure this process. The neutrino is extremely hard to observe, so in general, one sees an electron that all of a sudden gets a kick out of nowhere. In this case, it is a neutrino that has scattered off the electron by means of a Z-interaction.

The second diagram tells the same story, but just with different particles. Note that in both diagrams, the fermions can be both left-handed and right-handed. This is a consequence of the fact that the Z^0 is a mixture of the W^3 and B . So, any particle with hypercharge will feel the Z^0 to some extent, even when, like fermions with right-handed chirality, it does not feel the $W^{+/-}$.

B. Creation /annihilation interactions

A Z-boson can be produced when a fermion turns to its anti-self, i.e., to its anti-particle. This is an annihilation process (see Fig. 17.13). The Z-boson can decay again and create another pair, in this case a quark—anti-quark pair that forms a Φ meson.

In this diagram we also have fermions with a right-handed chirality that can be created and annihilated through a Z-boson. The Z-boson can decay into all fermion—anti-fermion pairs that exist. This has an interesting implication for the number of particle families, which we will discuss later in the section “family business”.

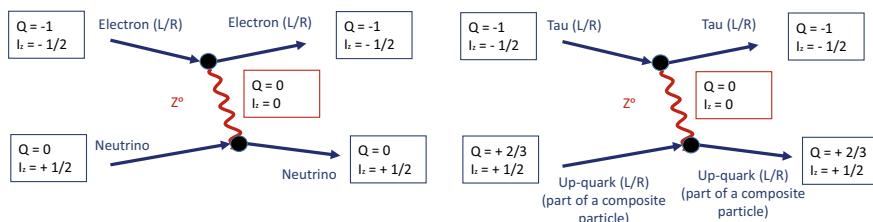


Fig. 17.12 Force produced by a neutral current process

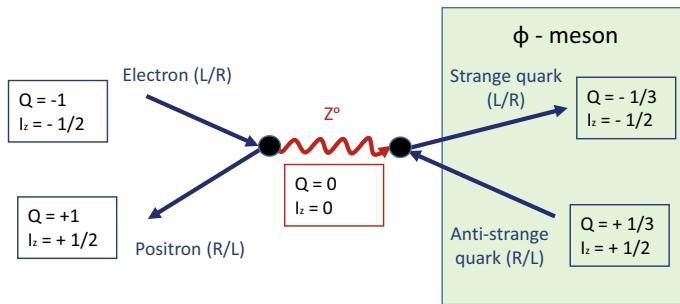


Fig. 17.13 Pair creation and annihilation via a Z° -boson

17.3.4 Z - W Self-interactions

Since the bosons of the weak interaction carry a weak (isospin) charge themselves, they can interact with each other. This type of interaction is called a self-interaction. Let's see what these interactions look like.

A. Creation/annihilation interactions

The W^+ and W^- can annihilate each other and produce a Z° (see Fig. 17.14). Put differently, they can turn into each other while producing a Z° . Similarly, the Z° can decay into a W^+ and a W^- . Consequently, we see that in fact W^+ and W^- are each other's anti-particles. The anti-particle of Z° is Z° itself, just as for the photon. A photon can also produce a W^+/W^- pair. When a Z° produces a W^+/W^- pair, this pair must be virtual and can only live for a very short time, given their considerable mass. The chances of them being absorbed by another particle are extremely low. Therefore, the W^+ and W^- will most likely get reabsorbed and become a Z° again. Consequently, this process is usually not considered when the decay possibilities of Z° are discussed.

B. Weak vacuum polarization

Just as an electron can produce a photon that splits into a charged particle and its anti-particle for a while, a fermion can produce a Z° in a similar fashion. The Z° can

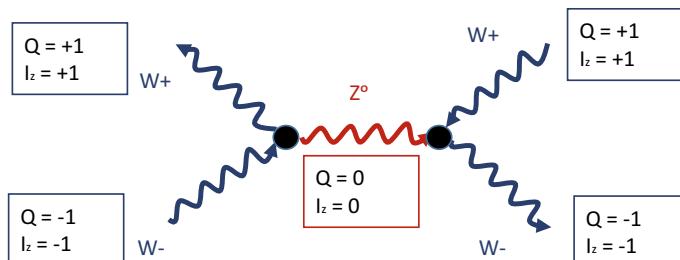


Fig. 17.14 Creating and annihilating W^+/W^-

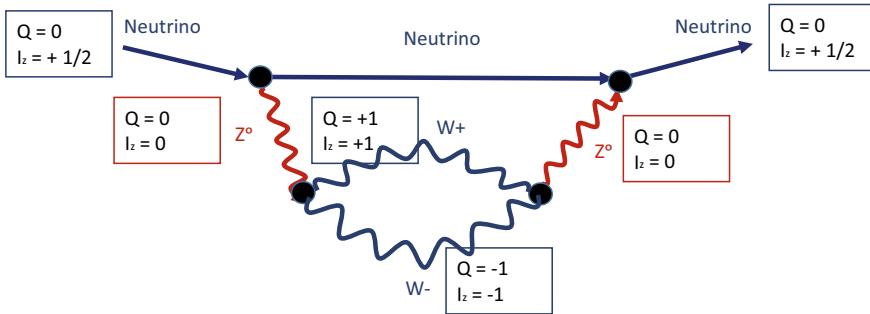


Fig. 17.15 Weak vacuum polarization

split into a W^+/W^- and this split can polarize the vacuum (see Fig. 17.15). So, in the presence of an electric field, the two charged disturbances influence the electric field and create some dielectric in the vacuum. But considering the masses of the W 's and the Z° , this process is relatively rare.

17.3.5 Neutron Decay

It is not only fundamental particles that change to a different face under the weak force. The same happens to composite particles such as the neutron. A neutron consists of three quarks: one up and two down. If we add up their charge, we get $+2/3$ (up) $- 1/3$ (down) $- 1/3$ (down) $= 0$. One of the quarks in the composite can change into its other face. So, a down quark can change into an up quark, while producing a W^- disturbance (see Fig. 17.16). The W^- can decay into an electron and an anti-neutrino. The changed quark has altered the composition of the neutron: it is now two up quarks and one down. Consequently, its charge is $+2/3$ (up) $+ 2/3$ (up) $- 1/3$ (down) $= +1$. The neutron has turned into a proton as a consequence of the rotation of one of its quarks. Hence, a neutron can decay into a proton, an electron, and an anti-neutrino. Note also that the total isospin of the neutron is $-1/2$, while that of the proton is $+1/2$. Consequently, the proton and the neutron can be considered as two faces of the same particle, but only by considering their constituent quarks. So, a neutron and a proton can be turned into each other under SU(2) (broken) symmetry and produce a W -boson.

The decay process is very slow due to the mass of the W 's. In other words, it is a rare process, so it takes a while (on average) before a neutron decays into a proton. A free neutron takes about 15 min to decay into a proton.

The neutron has a greater mass than the proton. This makes it possible for this type of decay to happen. The energy of the proton + electron + neutrino together is still less than the energy of the neutron. Consequently, the decay products take momentum out of the interaction. The other way around, a proton turning into a

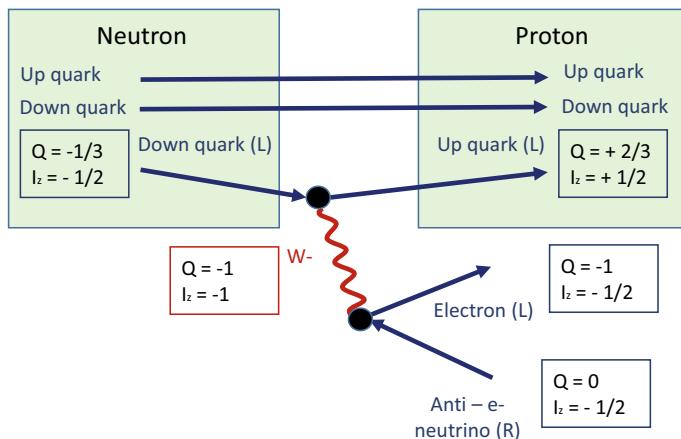


Fig. 17.16 Neutron decay

neutron is not possible, since the proton is less massive. It would require the capture of an energetic W-boson making an up quark change into a down quark. Since the weak interaction is already very short-range and W-bosons are not abundant, the chances for this to happen are astronomically small for a free proton.

17.3.6 Radioactive Decay

Neutron decay within a nucleus is called radioactive decay or β -decay. A nucleus consisting of N protons determines the chemical characteristics of the atom. N effectively determines the element. For example, six protons gives carbon. However, when a neutron in the nucleus decays into a proton, it changes the element. The nucleus will now have $N + 1$ protons and become a chemically different element. In the decay process an electron and neutrino are produced. For historical reasons the energetic electron produced this way is called β -radiation. Hence, β -decay. This is a type of ionizing radiation that is associated with radioactivity.

Carbon in its stable form has 6 protons and 6 neutrons. It is called C-12. C-13 exists as well. It has 7 neutrons along with the 6 protons that determine the atom to be carbon. It is much less abundant though. Of all carbon atoms, about 1% are C-13. Then there is C-14. It consists of 6 protons (still carbon!) and 8 neutrons. This form is not stable. One of its neutrons can decay into a proton. The result is a nucleus with 7 protons and 7 neutrons. An atom with 7 protons (and hence 7 electrons swirling around) is called nitrogen. So, C-14 decays into N-14. It takes about 5730 (± 40) years before half of a bunch of C-14 atoms is transformed into nitrogen. This is called its half-life.

This process is used on earth to “carbon-date” historic finds. Biological materials (such as wood) pick up C-14 from the atmosphere when they are alive. In the atmosphere a stable fraction of C-14 is available, since it is produced there. The production is in equilibrium with the decay and hence a stable amount can be found in the atmosphere. For example, when a tree dies, it stops picking up C-14 from the atmosphere. Consequently, the C-14 will decay and the amount of C-14 in the tree’s wood will decrease over time. So, by measuring the relative amount of C-14 in an old artefact made of wood, we can tell its age.

Not every neutron will decay within a nucleus. Whether or not it will do so depends on the stability of the nucleus. This is related to pion exchange between protons and neutrons, since this is what keeps protons and neutrons together in the nucleus. You might say that pion exchange is another way of changing protons into neutrons and vice versa. So, when this process is efficient, neutrons never exist very long. Hence, the probability of decay is very low. It would take us too far to get to the bottom of this, but in essence, the longer a neutron is a neutron in the nucleus the greater the probability of decay. Usually, the number of neutrons must be (significantly) higher than the number of protons to get a neutron to exist long enough as a neutron. Hence, the larger the surplus of neutrons in a nucleus, the less stable the nucleus gets (very generally speaking).

There is another type of β -decay. This is concerned with a proton changing into a neutron. Huh? That wasn’t possible, right? This process cannot happen to a free proton because it is lighter than a neutron. So where could we get the energy to make this happen? In a nucleus, the energy can come from the binding energy between the constituents of the nucleus. In that case, in a nucleus, a proton can decay into a neutron and produce a positron and a neutrino (see Fig. 17.17). The total binding energy of the resulting nucleus is equivalently lower than the total binding energy of the nucleus prior to the decay.

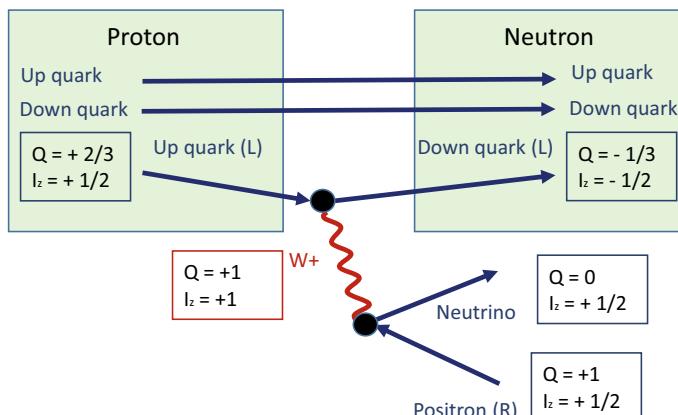


Fig. 17.17 β^+ decay of a proton, changing into a neutron

This process also depends on the stability of the nucleus. Proton decay requires the number of protons to be (significantly) higher than the number of neutrons. This enhances the probability for a proton to decay as a sitting duck. The higher the surplus of protons in a nucleus, the less stable it is (again very generally speaking).

The transmutation of the nucleus plays an important role in nuclear fusion processes in stars. If transmutation were not possible, the fusion process would not get far and the sun would not be able to power itself. Basically, the fusion process in the sun draws its energy from two steps: (1) putting nuclei together to form a new nucleus and (2) transmuting that nucleus into a stable version. Energy is released in these two steps and this powers the sun. Fusion can deliver energy until the formation of iron. From there on, heavier elements can only be created at the cost of energy. Hence, these elements are mostly formed in the end stages of a star's life, during an energetic process such as a supernova explosion. Consequently, most of these heavier elements are rarer than the lighter elements. For example, gold is much rarer than iron.

So, we can say that none of this (and none of us) could exist without the ability of particles to show their other face. The broken symmetry makes this process a high energy process as well as a less frequent one.

17.3.7 Cabibbo Rotation

So far, we have said that the fundamental left-handed fermions have two faces. They can change to their other face by producing a $W^{+/-}$ boson. This is a consequence of the SU(2) symmetry between the two faces. We have also seen that particles can mix. The mixing of W^3 and B into the photon and the Z^0 are examples.

It turns out that a quark can turn into a quark of a different family. Does this mean that quarks have more than two faces to change to? How could that happen? One assumes that the quarks in the three families are mixed somehow as a consequence of the symmetry breaking. Before symmetry breaking, a quark of one family could not turn into a quark of another family. One consequence of symmetry breaking is that the quarks gain mass, and they each get a different mass. Moreover, the massive $W^{+/-}$ bosons that propagate the change are different from before the symmetry breaking. So, it may not be surprising that this alters the way such a change works, between the two faces of a quark.

So how does quark mixing create the possibility of one quark turning into a quark of another family? Take for instance a down quark. It mixes a little bit with the strange quark and the bottom quark. That means that the wave of a new down quark (i.e., after symmetry breaking) looks like that of an old down quark and a little bit like that of a strange quark and a bottom quark. The same goes for e.g., a strange quark. The wave of the new strange quark looks a little like the wave of a down quark. So, when a new strange quark changes face, it may sometimes do so right at the time that it looks like a down quark. Consequently, it turns into an up quark, the other face of a

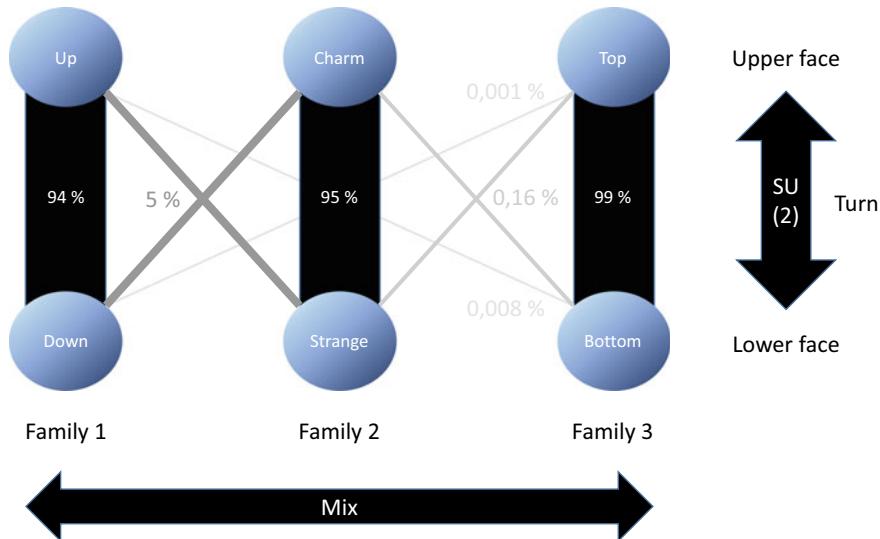


Fig. 17.18 Probabilities of transformations between the quark families

down quark. Hence, the mixing creates the possibility of turning into the other quark faces of all three families.

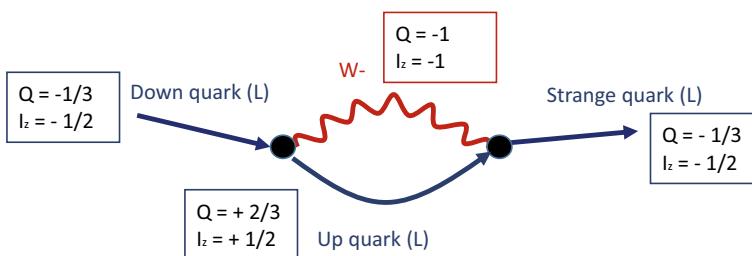
One mix takes place between the down, strange, and bottom quarks. These are all “lower face” quarks in Fig. 17.18. The other mix takes place between the up, charm, and top quarks. These are all “upper face” quarks in Fig. 17.18. So lower face quarks do not mix with upper face quarks. This means that the two faces remain strictly separated and it takes an isospin SU(2) rotation to get from one face to the other. It is just that the families are no longer strictly separated. The isospin SU(2) symmetry still exists. In fact, strictly speaking, the symmetry hasn’t changed. A down quark can only turn into an up quark. It’s just that quarks of the same face in different families get mixed a little.

The mixing amplitudes between the families are generally small. They have been measured [Ref. 20]. Table 17.4 gives the results as far as they have been measured at this point in time. For example, the new down has an amplitude of 0.225 to be a strange quark. As a strange quark, a rotation can transform it into a charm quark. The probability that this will take place is proportional to the amplitude² to be a strange quark to begin with, so in this case $(0.225)^2 = 0.05 = 5\%$.

Table 17.4 shows that it is still much more likely that a quark will turn into its other face within its own family (bold percentages). However, it is no longer impossible to turn into the other face that belongs to another family. In particular, the new strange quark has a 5% probability of turning into an up quark. Equivalently, the new down quark has the same probability of turning into a charm quark. So, the probability of turning into the other face of another family is crosswise equal (see Fig. 17.18).

Table 17.4 Mixing amplitudes between the quark families

New quark	Mixing amplitude	Chance of changing to up (%)	Chance of changing to charm (%)	Chance of changing to top (%)
New down	0.97 down + 0.225 strange + 0.0035 bottom	94	5	0.0012
New strange	0.225 down + 0.973 strange + 0.04 bottom	5	95	0.16
New bottom	0.0087 down + 0.04 strange + 0.999 bottom	0.0076	0.16	99

**Fig. 17.19** Cabibbo rotation of a down quark into a strange quark

The transformation of a quark into a quark of a different family was first described by Nicola Cabibbo [Ref. 71]. He was the theoretical physicist who first solved the puzzle of the weak decay of particles containing a strange quark. The weak mixing angle θ between the quarks is called the Cabibbo angle [Ref. 23, handout 12]. This is another way of describing how much quarks look alike. Figure 17.19 shows an example of a Cabibbo rotation.

In Fig. 17.19 a down quark turns into an up quark and produces a W-disturbance. This recombines with the up quark, which turns into a strange quark. So, we see that at every step of the process a quark turns some way it is allowed to turn (from lower face to upper face and back). But the end result is that a down quark has become a strange quark. This process requires energy, as the strange quark is more massive than the down quark.

17.3.8 Neutrino Oscillations

Neutrino oscillation was first predicted in 1957 by Bruno Pontecorvo. Since then, it has been observed in a variety of experiments [Ref. 85]. Neutrino oscillation is an

important mechanism in the explanation of the solar neutrino deficit. The standard model and solar models together predict the number of neutrinos that would have to be produced by the sun as a consequence of its nuclear processes. However, measurements of solar radiation find a large deficit in the number of neutrinos produced. The sun produces neutrinos of the electron type in its nuclear processes. One way to explain why there are not enough electron neutrinos coming from the sun is to assume that electron neutrinos are not always electron neutrinos, but sometimes behave as muon neutrinos or tau neutrinos. Consequently, there are enough neutrinos coming from the sun, but they are not always detectable as electron neutrinos. This assumption is in line with other experiments using a variety of detectors that show such oscillation between neutrino flavours. In 2001, the Sudbury Neutrino Observatory proved from its experimental data that neutrino flavour oscillation is the cause of the solar neutrino deficit.

So how do neutrino oscillations work? And why do they exist [Refs. 85, 86]? First of all, to be able to oscillate between the three flavours, a neutrino needs to have the possibility to “switch”. That can only happen when there is some overlap between the wave functions of the three flavours. So, the three neutrino types must look a lot like each other. Suppose the three neutrino types were exactly the same. Then they would become indistinguishable. If there were a very small difference, they would be able to look like one another for a long time. The bigger the difference, the smaller the probability of a switch and consequently, the shorter it could impersonate another flavour. So, the oscillation time depends on the difference between the neutrinos. The bigger the difference, the shorter the oscillation time.

So, what could be that difference? In comparison, the electron, muon, and tau differ only by their mass and flavour. But if their masses were (almost) the same, they might oscillate between their flavours as well. On the face of things, neutrinos seem to have the same properties, such as zero charge, all being left-handed, and connecting to the weak force in the same way. The only way they could differ slightly would be if they had a small mass that differed between flavours. The mass difference would then be directly related to the oscillation frequency. How can we see that?

A difference in mass constitutes a difference in energy. So, we may use the uncertainty relations (see Sect. 4.4). For a certain amount of time, an extra amount of energy can be available. So, a neutrino may switch into a neutrino of a heavier type for a time Δt , depending on the amount of energy ΔE required to cover the extra mass. The greater the difference in mass, the shorter the oscillation time. So, by measuring the oscillation time, we can create an upper limit for the mass difference between the neutrino flavours. In any case, this requires neutrinos to have mass.

Neutrinos are produced as a particular flavour in a weak interaction process, in the eigenstate of that flavour. But as a consequence of the oscillation process, they do not stay in that flavour eigenstate. Consequently, they travel in a mass eigenstate which is a mix of the three neutrino masses, while switching continuously between the flavours (see Fig. 17.20). The difference in mass leads to a difference in (mass) frequency of the mass eigenstate. Remember, frequency is proportional to energy (hence mass). Depicting this as three waves with different (mass) frequencies that are in a superposition with each other, we can understand that the wave for flavour

Flavour 2 and 3 annihilate each other so that flavour 1 is the one measured. This stays the case as long as flavour 2 and 3 are in opposite phase. That time can be long when the frequency (wavelength in time) difference is small between 1, 2 and 3.

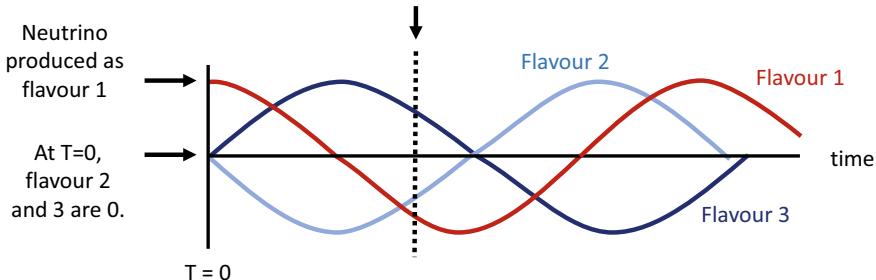


Fig. 17.20 Three waves for the three neutrino flavours. Their mass difference translates into the different frequencies of the three waves. Since the x-axis represents time, the three different frequencies translate into three different wavelengths in a time dimension. That way we can show how, at some times, flavours 2 and 3 will cancel each other out and measurement will reveal flavour 1. After a while, the waves will have shifted due to their difference in frequency (hence period in time) and flavour 1 might cancel out against flavour 2, leaving flavour 3 with the highest probability of being measured

1 will sometimes be maximal, while the waves for flavours 2 and 3 cancel each other out. After a while, waves 1 and 2 might cancel each other out and measurement will reveal flavour 3. This will thus continuously change the probability of measuring flavours 1, 2, or 3. The smaller the difference in mass between 1, 2, and 3, the smaller the difference in frequency and the longer the oscillation period. Hence, the longer a neutrino will be one type of neutrino before it switches to another.

This oscillation process can be compared to three coupled harmonic oscillators (see Sect. 8.2.1). There can be three modes in which the neutrinos may oscillate between flavours. Just as with Cabibbo rotation, we can define a mixing angle between the three flavours. Just keep in mind that Cabibbo rotation was a different process that caused quark flavours to mix. For the neutrinos, the mixing angles will indicate the mass difference between the neutrinos.

So, what is the origin of the neutrino mass? The bare mass of a particle is generated by connecting to the Higgs field. However, we will see in Sect. 17.4 that the Higgs field makes massive particles switch between left- and right-handedness. The problem is that the neutrino only comes in a left-handed version. So, Higgs cannot be responsible for the neutrino mass. However, there can be other sources of mass [Ref. 21]. Maybe the right-handed version does exist, but it does not interact with any of the three forces. Hence, we would not be able to detect it, even though it is still there. Alternatively, there may be no right-handed neutrino and the neutrino could get its mass through an indirect effect involving Higgs. This would actually be an interesting option since it would explain why the neutrino mass is so small. Basically, any potential can be responsible for creating mass, as we have seen before. As a matter of fact, the standard model does not provide an answer to this problem. At present, science is

searching for the right mechanism to be added to the standard model to explain the neutrino mass.

17.3.9 Concluding

Looking at the transactions we have discussed, the isospin SU(2) symmetry clearly still exists. But now the charges of the composites of each pair are no longer the same. The change can still occur since the W's can carry U(1) charge and hence propagate the difference away. Moreover, the SU(2) symmetry is no longer perfect. The quarks turn out to be mixed, so there is a certain probability of turning into the other face of a different family.

Symmetry breaking has had an interesting effect on both the U(1) and SU(2) symmetry and has left us with a new charge and new possibilities for particles to turn into each other. The resulting transaction types play an important role as an energy source for the sun, as well as the long sought alchemy that can change everything into gold; or rather, create all the elements we know in nature.

17.4 Fermions Gaining Mass

We saw before that fermions gain mass due to the interaction between their field and the Higgs field, which is non-zero everywhere. We also saw before that the interaction with Higgs can cause the chirality of a fermion to flip. So, the Higgs potential would switch left-handed fermions to right-handed ones and back all the time. The mass potential as caused by Higgs depends on this process. When this process does not happen, the mass potential is 0 [Ref. 8, p. 439].

This switch is different than the SU(2) rotation. It is not born from a symmetry (that creates a gauge field), but from an interaction with another field (Higgs). Since it does not concern a symmetry, the left-handed and right-handed electron are not two faces of the same particle. They are different particles, but Higgs switches them together, and this may work by precession, as described in the chirality section (see Sect. 13.4).

This switching process happens very fast. Once the electron is flipped to its other chirality, it interacts with the constant Higgs field again and flips a second time. Therefore, the frequency of flipping depends on the strength of the coupling to the Higgs field. The frequency of switching also inhibits an energy, as there is energy in a precessional motion. It is the energy we must provide to an electron excitation so that it can be a real, steady quantum. As soon as there is any disturbance of the electron field, it will start to interact with the Higgs field and it will start flipping. If there is not enough energy in the electron disturbance, the potential of the Higgs field will damp the wave and the disturbance will not become a full excitation. So,

the energy provided must be at least enough to get the flipping going in the Higgs potential.

Summarizing, there are two ways to view how an electron becomes massive:

- The Higgs potential is like a spring that changes the characteristics of the wave. The springs make it more difficult to start a wave in the electron field. Hence, more energy must be put in, equal to the mass of the electron.
- The Higgs potential flips the electron at a high frequency. That frequency represents an energy equal to the mass energy of the electron. Starting an excitation in the electron field means being able to get the flipping process started, and that requires that mass energy.

The two descriptions look like different ways to view the same situation, but they are in essence the same. The potential creates the precession and this combination determines the characteristics of the spring that influences the wave. Both flipping and springs are a way to store energy, which is the rest-mass energy of the electron.

The switching between right- and left-handed happens so fast that it seems as if the left- and right-handed electron are one particle. They are in fact two particles, but that is not how they appear to us. For example, the very heavy top quark switches from left to right and back at a staggering frequency of 10^{26} times per second [Ref. 21]. That is a 1 with 26 zeros switches and back per second. So, if you think you have a left top quark, soon you will have a right top quark. And before the blink of an eye, it will have flipped a billion times. So clearly, the top quark as it appears to us is an average of a left-handed and right-handed top quark. An average yes, but we will never have both at the same time! Hence, they are not a composite or two faces of the same particle. They just happen to switch really fast.

The coupling between the electron and Higgs is much lower. So, the electron flips much less frequently between left and right. In fact, only about 0.000003 times as frequently as for the top quark [Ref. 21], and so its mass is also 0.000003 times that of the top quark. This is the way Higgs “gives” mass to the excitations of all fermion fields (except the neutrino). Strictly speaking, it is not Higgs that “gives” mass to the excitations of a fermion field. Higgs ties the fields to a potential that starts to count when the field is excited (see Fig. 17.21). Hence, all Higgs does is require *you* to

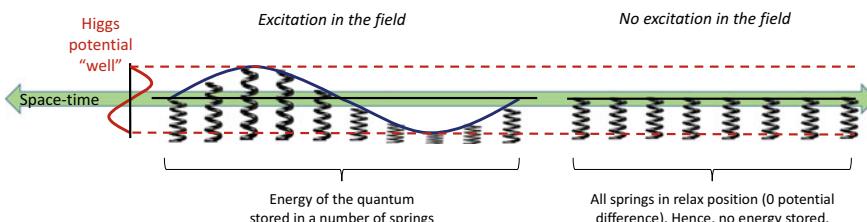


Fig. 17.21 The Higgs potential starts to count when a field gets excited. In space and time, the Higgs field is constant. But when a field gets excited, the connected Higgs potential starts to provide a potential difference

put in enough energy in an excitation. So, it's actually *you* that gives mass to the excitation when you create it, not Higgs!

If we could take out the Higgs field, the particles would stop switching and they would become massless left-handed and right-handed particles again.

We saw that the neutrino does not couple to Higgs because it has no right-handed version. Hence, it cannot flip and cannot gain mass this way. It's as though the non-zero Higgs field only wants to couple to excitations that have left- and right-handed cousins.

So, a particle's (bare!) mass is directly related to the coupling strength of its field to Higgs. Considering the huge variations in particle masses, we must conclude that the coupling strength to Higgs varies just as much. That is interesting because the U(1) and SU(2) forces (and SU(3) as we will see) have “quantized” coupling strengths that do not differ much between particles. The coupling to Higgs, by contrast, is not quantized at all and differs by a factor of over 300,000 between the electron and the top quark. The reason could be that Higgs does not seem to originate from a symmetry, while the force carrying bosons do. Otherwise, we still have no answer to this problem.

Is all mass caused by Higgs? No. We saw before that any potential can give mass to waves. They are also like springs attached to the wave. We saw before that potential energy (binding energy) gives mass as well. Especially when we look at the strong force, the binding energy is very large. For example, within a proton or neutron, the kinetic energy of the quarks jiggling about as well as the binding energy of the quarks inside the proton or neutron add up to approximately 99% of the mass of the proton or neutron. So, all types of energy add to the mass, and they can be substantial.

17.5 Parity Violation and CPT Symmetry

We saw before that the weak interaction does not treat the vacuum in a symmetric way. This fact is called parity violation. “Violation” means that the weak force violates the symmetry between a quantum and its parity reverse (its other chirality in the case of fermions) by treating it differently.

This violation of parity has been observed. The first measurement that showed such a violation of parity was made in 1956 by Chien-Shiung Wu et al. [Ref. 62]. The idea of the experiment was to see whether the electrons coming from beta-decay in cobalt-60 would favour a direction relative to the nuclear spin. So how would that show that parity is violated?

Basically, when beta-decay occurs, a left-handed down quark turns into a left-handed up quark by emitting a W^- boson. This decays into an electron and an anti-neutrino (see Fig. 17.16). The W^- boson and hence also the electron are emitted in a certain direction relative to the nuclear spin. However, if the down quark happened to be right-handed, it would turn into a right-handed up quark while emitting the W^- (and subsequently the electron) *in the opposite direction!* Note that in those days it was not known that right-handed quanta do not feel the weak force. The opposite

direction follows from the fact that the right-handed quark is a parity-reversed version of the left-handed quark. Hence, all directions have been reversed, including that of the emitted W^- .

In the experiment, the cobalt would change into nickel under beta-decay. However, the nickel would be in an excited state. It would fall back to its ground state under emission of two photons. The direction of these photons depends on the direction of the nuclear spin. The electromagnetic force is known to conserve parity. Hence, it makes no difference to the photons whether they originate from a left-handed or a right-handed process. They would equally likely be emitted after a left-handed process as after a right-handed process. Furthermore, the quark that turned into its other face will have swapped chirality many times before the photons get emitted to take the nickel back to its ground state. Therefore, the photon emission process will not even know about the history of the β -decay. Consequently, the number of photons going in one direction would tell us simply how many atoms are spinning one way and how many atoms are spinning the other way. In the experiment this was 60–40. So, 60% were spinning one way and 40% were spinning the other way.

Now suppose that the weak interaction does not distinguish between left-handed and right-handed quarks. The probability that they decay while emitting a W^- is the same. Then the distribution of directions the electron comes out with will be 50–50. Why is that? Suppose we have 60% of the atoms spinning in one direction. The left-handed decays will emit electrons in direction 1 and the right-handed decays will emit electrons in direction 2. The quarks involved swap their handedness continuously and very fast, so there is on average a 50% probability of decaying in left-handed mode and a 50% probability of decaying in right-handed mode. So, from the atoms spinning in one direction, 30% come out in direction 1 and 30% in direction 2. Now we take the 40% atoms spinning in the other direction. The left-handed decays will now emit electrons in direction 2 and the right-handed decays will emit them in direction 1. Again 50–50, so 20% go in direction 1 and 20% in direction 2. So, in the end there are $30 + 20\%$ electrons emitted in direction 1 and $30 + 20\%$ in direction 2. Hence, 50–50 (see Fig. 17.22).

Now suppose that the weak interaction only takes place in *right*-handed quarks. What then? From the 60% atoms spinning in one direction 30% come out in direction 1 and none in direction 2. From the 40% atoms spinning in the other direction, 20% come out in direction 2 and none in direction 1. So now the distribution of electrons coming out is not 50–50, but 60–40. It looks exactly like the distribution of the photons coming out, equal to the relative number of atoms spinning one way. If the weak interaction only takes place in *left*-handed quarks, the same result will appear, but opposite: 40–60 (see Fig. 17.22).

The result of the experiment was that the distribution of electrons was the same as that of the photons, but in the opposite direction [Ref. 62]. Hence, the conclusion of the experiment was that weak interactions only take place in left-handed quarks.

This is strange: a direction or handedness is favoured in nature. This seems very counter-intuitive. In relativity we stress the fact that space–time is symmetric. There should be no law of nature that is able to distinguish between directions or velocities.

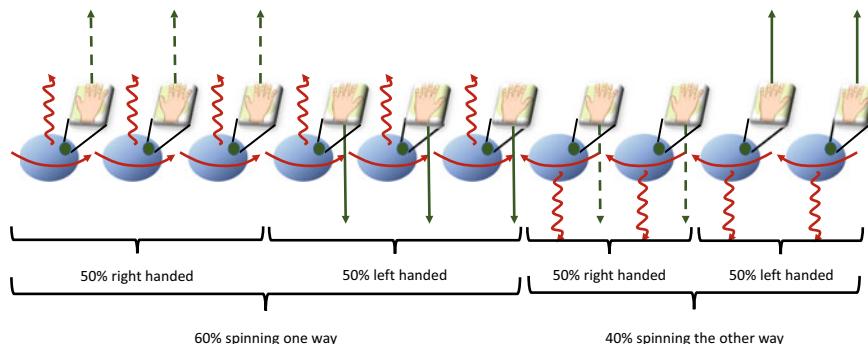


Fig. 17.22 Parity violation. If there were no difference between left- and right-handed weak decay, there would be in total 5 green arrows going up and 5 arrows going down. If only left-handed quarks decayed via a weak interaction, only the solid green arrows would count (the left-handed ones) and we would find three going down and 2 going up. This is the same ratio as the number of photons (red) going up (6) versus going down (4), but in the other direction

But now we see that there is actually an absolute sense of handedness in nature. So, does this violate relativity? Well, read on.

Let's look at some other symmetries. For instance, charge conjugation. This is called C-symmetry. Charge conjugation is nothing else than swapping the charge of a quantum. By charge we mean not only electromagnetic charge, but also the charges associated with all other forces, including isospin and colour charge. If we swap the charges of all quanta in a system, the electromagnetic interaction will treat the system in exactly the same way. So, this is a symmetry for the electromagnetic laws of nature. How about the weak force?

When we swap all charges of, e.g., a left-handed electron ($Q = -1; I_z = -1/2$), we get a left-handed electron with a positive electromagnetic charge (+1) and a positive isospin charge (+½). This looks exactly like a left-handed positron. So, the charge conjugate of a left-handed electron would look like its left-handed anti-version. It turns out that the weak force does *not* couple to left-handed positrons! This goes for all particles, so we say that the weak force maximally violates charge conjugation.

Thinking about this suggests another idea. What if we swap charge (C) *and* we swap parity (P) at the same time? This combination of the two symmetries is called CP symmetry. After a CP swap, a left-handed electron would become a *right*-handed “positron”. It is not yet a full positron, since time has not been swapped. Nevertheless, the weak force would couple to a right-handed positron, so this combined CP symmetry looks as though it may be alright. But is it?

In 1964 an experiment was performed by James Cronin, Val Fitch, and coworkers [Ref. 63] using kaon decay that showed that even CP symmetry could be violated.

A neutral kaon consists of a strange quark and an anti-down quark. The neutral anti-kaon consists of a down quark and an anti-strange quark. The two kaons can oscillate between each other through weak interactions (see Fig. 17.23). This means that they are effectively indistinguishable. Consequently, we need to consider their

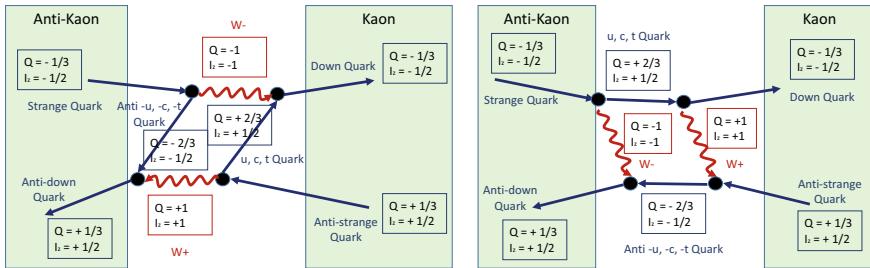


Fig. 17.23 The two ways for a kaon to oscillate with its anti-particle

superposition (just like we did when we mixed W_3 and B). We can have a positive superposition and a negative one. The two superpositions decay in different ways. The negative superposition is called $K(\text{short})$ since it decays quickly into *two* pions. The positive superposition is called $K(\text{long})$ since it takes much longer to decay. It decays into *three* pions.

Both superpositions turn out to be CP eigenstates. This means that when we apply a charge conjugation and a parity operation, we get the same state back multiplied by a constant. Applying CP to both superpositions gives:

$$\text{CP}[K(\text{short})] = +K(\text{short})$$

$$\text{CP}[K(\text{long})] = -K(\text{long})$$

Therefore, $K(\text{short})$ is called a CP-even eigenstate and $K(\text{long})$ a CP-odd eigenstate. The CP operation produces the same state or minus the same state. Hence, these states show a clear symmetry under CP operation.

Experiments with kaons showed that a very small portion of the $K(\text{long})$ eigenstates decayed like a $K(\text{short})$ kaon into *two* pions. Consequently, this process violates the above CP symmetry. The level of CP violation is very small. About 0.2% of the decays of $K(\text{long})$ occur via two pions.

One way to explain the observed CP violation is by considering $K(\text{long})$ and $K(\text{short})$ as a slightly different mix of the neutral kaons. This can be described by remixing the CP eigenstates into states that have (1) a lot of $K(\text{long})$ and a little of $K(\text{short})$ and (2) a little of $K(\text{long})$ and a lot of $K(\text{short})$. Consequently, the new $K(\text{long})$ can decay into two pions for a small fraction of the decays.

Other decay processes of kaons occur through leptons. The kaon decays into π^- , e^+ , and a neutrino and the anti-kaon decays into π^+ , e^- , and an anti-neutrino (see Fig. 17.24).

When we observe the decay of the $K(\text{long})$ combination, it turns out that the anti-kaon decay is 0.7% more likely than the kaon decay [Ref. 23, handout 12]. This again implies that the $K(\text{long})$ mix is not an even mix of a kaon and an anti-kaon. Another consequence is that this shows a clear difference between matter and anti-matter.

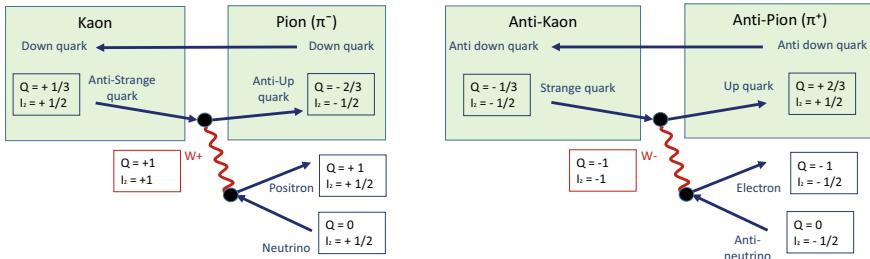


Fig. 17.24 Some decay modes of the kaon and anti-kaon

Considering that kaons and anti-kaons oscillate continuously between each other, electrons will come out more often than positrons.

The difference in the mix for K(long) and K(short) also implies that, in the oscillation between kaons and anti-kaons, the probability of going from kaon to anti-kaon is slightly different from the probability of going from anti-kaon to kaon. This is important as an anti-kaon is supposed to be a kaon that travels back in time. So, we might say that there is a difference in time symmetry as well.

Let's look at time-symmetry. This is a symmetry where we would see the same process when we reverse time. When we discussed the arrow of time, we saw that this is true for single interactions (except for the kaon, of course). In a large number of interactions, statistics come into play, as discussed before. When we reverse time in a *single* electromagnetic interaction, the process is exactly alike. So, electromagnetism is time-symmetric. Above, we saw that the weak interaction is not time symmetric as a consequence of its not being CP symmetric.

So how about CPT symmetry? A CPT operation is a charge conjugation, a parity operation, and a time-reversal. When we do this, we actually turn quanta into anti-quanta that have opposite chirality and opposite charge and travel in the opposite time direction. Hence, we enter into an anti-matter world. So, the anti-matter world should be an exact mirror world in which the laws of physics are the same. CP violation differentiates between matter and anti-matter, but CPT symmetry does not. When CPT symmetry holds, and a system violates CP symmetry, it must also violate T symmetry. The T violation must exactly cancel the CP violation in order to keep CPT symmetry intact. The weak interaction has been shown to do exactly this: it violates CP symmetry as well as T symmetry but leaves CPT symmetry untouched. We can see this from the kaon decay: the difference in decay between kaons and anti-kaons is exactly the same as the difference in the mix of kaons and anti-kaons in K(long). That mix is responsible for the CP violation.

In the section on the time order of events we saw that time reversal actually automatically reverses charge. So, does time reversal naturally include charge conjugation? When changing t for $-t$ in equations, the charge has to be reversed in the equations separately. Time reversal does not take care of a parity reversal, so this has to be included as well. Hence, in equations we have to *explicitly* swap time, charge (including electromagnetic, isospin, and colour charge), and chirality. After we have

done that, all matter has changed into anti-matter with the opposite electric, weak and strong charge, and opposite chirality (left/right-handedness). Recall that, when we described how charge reversal in a wave is a consequence of time reversal (see Fig. 10.7), we implicitly included parity reversal (i.e., the wave going in the other direction).

Interpreting charge conjugation as an “automatic gift” from time-reversal, we see that CPT symmetry effectively changes into parity and time reversal. This is a change of direction in space *and* time. This looks rather logical and familiar. When we change all directions (*including time*), everything should stay the same. So, let’s go back to our question about relativity: is that still violated now that we have included time in the symmetry?

In physics we say that CPT symmetry “commutes” with the total energy. This means that when we measure the total energy of a system and then carry out a full CPT transformation, the result is the same as if we first performed a CPT transformation and then measured the total energy. Put differently, the total energy of the system does not change as a consequence of a CPT transformation. This also means that CPT as a symmetry leads to a “conservation law”. It has been proven that CPT symmetry implies and/or is a consequence of Lorentz invariance. This is the so-called CPT theorem [Ref. 8, p. 139; Ref. 30, p. 201]. So, any system that violates CPT symmetry, also violates Lorentz invariance.

Lorentz invariance is what lies behind the theory of relativity. Each law of physics has to obey Lorentz invariance. This means that when we change from one frame of reference (velocity) to another the law must still be the same. So, we now see how to join the dots: CPT invariance is basically a symmetry under swapping all directions in space–time while its “conservation law”, Lorentz invariance, states that the laws of physics remain the same when we change frame of reference in space–time.

Another way to see this is by considering once again that the order of events is different in a different frame of reference (different velocity). As we saw before, this means that a particle in one frame of reference can be an anti-particle in another. An anti-particle is also a full CPT transformation of a particle. When CPT symmetry states that the laws of physics must be the same under a CPT transformation, it means that they must be the same for particles as for anti-particles. That is the same as saying that the laws of physics must be the same in one frame of reference (where the particle appears as a particle) as in the other frame of reference (where the particle appears as an anti-particle), which is Lorentz invariance.

In all experiments to date, CPT symmetry has not been violated.

Apart from kaon decay, more evidence has been found of (small) CP violations, e.g., in B-meson decay. Other possibilities have been theorized and some have been measured. As we saw before, such CP violation could be responsible for an imbalance in the amount of matter versus anti-matter, since CP violation distinguishes between the two types of matter. We observe in the current universe that there seems to be a surplus of matter over anti-matter. The question arises as to whether CP violation could be responsible for this imbalance. However, all types of CP violations together are not enough to explain the imbalance we observe in the universe. Hence, there

must be another (additional?) effect responsible for it. It remains interesting that CP should be so close to being a symmetry, since violations are so small.

Consequences of CPT symmetry are, among others:

- The mass and lifetime of particles must be the same as the mass and lifetime of their anti-particle counterparts.
- It can be shown that fermion waves change sign and bosons do not under two subsequent CPT operations. Applying two CPT operations one after the other is the same as a 360° rotation. Hence, this way it can be proven that fermions require two full turns (720°) to get back to their original state while bosons require just one (360°) [Ref. 72], showing once again the relation between relativity, the structure of space-time, and spin!

17.6 Family Business

So far, we have grouped fundamental particles into three particle families. Why is that? What makes them a family? There are three relations between the fundamental particles by which we can distinguish one family from another.

The first is SU(2). We know that the up and down quarks are two faces of the same particle. Symmetry breaking has caused other families to be able to turn into up and down quarks as well, but the probability for that is much lower. So, we can consider particles that turn into each other via the *main* SU(2) rotation to be members of the same family. This goes for the three groups of quarks and their anti-quarks. It also works for the three groups of leptons and their anti-leptons.

The second is chirality. The two chiralities of a particle are just each other's mirror image (plus a turn). We have seen that the Higgs field spirals them around so that they are left-handed at one moment and right-handed at the next. This "precession" is a consequence of the Higgs potential that causes the particles to gain mass. They consequently have the same mass. They also share the same charge, which means that they produce phase changes in the same way. Hence, they are linked together and bound by the same properties with the exception of the SU(2) symmetry. So right-handed particles are in the same family as their left-handed cousins.

The third is mass. There is a pair of light quarks, a pair of medium heavy quarks, and a pair of heavy quarks. Although none of the individual quarks have the same mass, the difference between the pairs is large. The same goes for the leptons. The electron is light, the muon medium heavy, and the tau is very heavy (about twice the mass of a proton). So, when we order them according to their mass, we know what particles belong to what family.

Hence, the first family consists of the lightest quark SU(2) pair: up and down, their left- and right-handed versions; and the lightest lepton SU(2) pair: the electron and electron-neutrino, the right-handed version of the electron; and finally, the anti-particles of all these.

These are also the most common particles. All the matter we meet in our everyday lives is built up from them. Why is it not built from any of the heavier families? Basically, the family members of the heavier families can all decay into the lightest versions and do so quickly. But the lightest family cannot decay further, so its members have a much longer lifetime. Consequently, the only place where we find the heavier families is in high energy processes, e.g., in particle accelerators, supernova explosions, and cosmic radiation hitting our atmosphere at high altitude.

Why are there only three families? We do not know. Measurements indicate that there are three, but accelerators keep looking for a further family. One reason we believe that there are only three is because of the decay of the Z° -boson. Z° can decay into all families [Ref. 17; Ref. 23, handout 14]. It decays into a particle and its anti-particle. Hence, the Z° -boson can decay in 24 ways:

- In 10% of the cases, it decays into one of the three leptons and their anti-particles
 - In 20% of the cases, it decays into one of the three neutrinos and their anti-particles
 - In 70% of the cases, it decays into one of the 6×3 quarks and their anti-particles.
- The number 6×3 comes from the fact that there are 6 quark flavours and 3 quark colours.

The neutrinos cannot be measured. We only know that the Z° has decayed into neutrinos when it disappears, giving away its energy and momentum. In this case, that energy and momentum went into neutrinos.

However, by measuring the other decay products, we know how often it should decay into neutrinos and how much energy goes into them. If there are two, three, four, or more families, the number of decay possibilities goes up. Four families would give the Z° 32 decay possibilities instead of 24. This would also mean that Z° would have a shorter lifetime. After all, the more decay possibilities there are, the faster Z° will decay into one of them. The shorter the life of Z° , the more uncertain its energy (see Fig. 17.25). So, when we measure the width (uncertainty) of its energy peak, we have an indication of the number of families there are that Z° could decay into.

This can also be viewed in the following way. When Z° decays into a particular particle pair, that decay causes a width in the energy peak of Z° . If Z° did not decay at all, its energy peak would be very sharp since Z° would live for a very long time. If it decayed into one particle pair, the energy peak would be a little less sharp. So, the total width of the energy peak of the Z° must be a sum of the energy widths caused by each pair it decays into.

You might wonder whether a fourth family might be so heavy that the Z° would not be able to decay into those particles. It might simply lack the energy to produce such a pair. However, it might do so temporarily, hence producing a *virtual* pair which decays into lighter particles with an energy that does add up to the energy of the original Z° . This would be a less common way to go, and hence would not add much to the energy width of the Z° . However, Z° would at least be able to decay into the neutrino of the fourth family, assuming its mass to be low (like all neutrinos so far). Hence, the decay width must be greater when there is a fourth family.

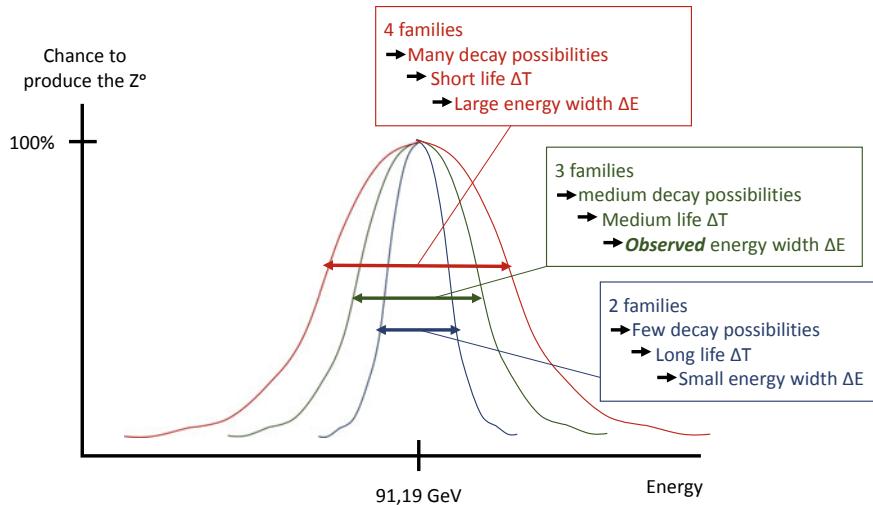


Fig. 17.25 How the width of the energy spectrum of Z^0 tells us how many families there are

CERN has measured the width of the energy peak of Z^0 [Ref. 73]. This width is about 2.5 GeV and it is in excellent agreement with the number of decay possibilities given by three families of particles.

In conclusion, we can say that there are three families of particles. We can organize the known fundamental particles into these three families based on their mass and their main SU(2) relation. The right-handed particles are connected to them through Higgs and share most properties. The anti-particles of all of these must then be added to complete the family.

17.7 Fundamental Particle Overview 3

We are now ready to put all fermions into a new particle overview (see Fig. 17.26). This overview contains the situation after symmetry breaking and includes U(1) and SU(2) relations. The grey-shaded region shows the U(1) symmetry propagated by the photon. After symmetry breaking, the neutrinos have zero charge. Hence, the grey region is limited to the other particles. Hypercharge no longer plays a role, as the particles no longer connect to the B-boson.

SU(2) still gives the same transformation between the two faces of each particle. However, the connection between the two faces is thinner due to the fact that there are other ways to turn. The other ways to turn are indicated by the blue arrows. These are thinner when the probability of the corresponding change is lower. The arrow points in the direction of the lighter particle since it will be easier to change from heavy to light than the other way around.

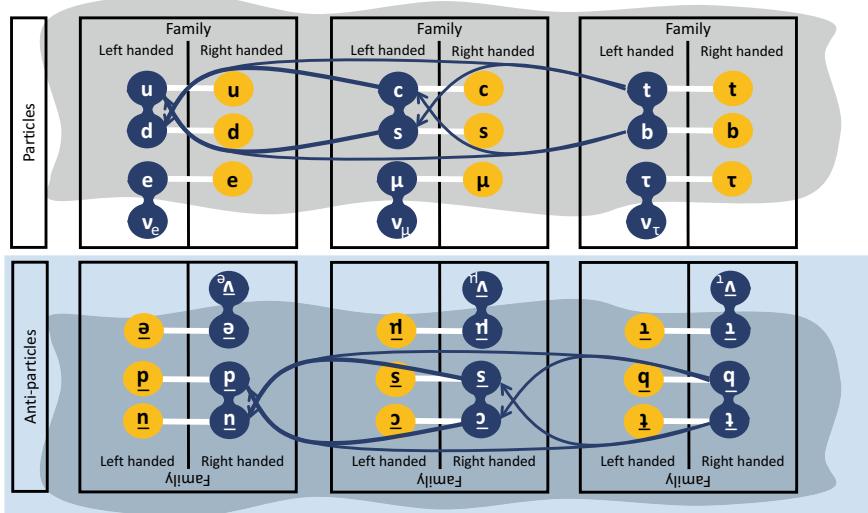


Fig. 17.26 Fundamental fermion overview after symmetry breaking, excluding the strong interaction. The U(1) symmetry is the region shaded grey. The particles in that region produce and feel phase shifts. The blue particles experience SU(2) symmetry and are able to turn into the particles that are connected to them by a blue line. The width of the line indicates the probability for that to happen

With this overview we finish our chapter on the weak interaction. The next step is to look at the strong interaction. The theory that concerns the strong interaction is called “Quantum Chromo Dynamics” or QCD for short.

Chapter 18

The Strong Force: Quantum Chromodynamics



Quantum chromodynamics (QCD) is the theory that describes the strong force between quarks. Just as quantum electrodynamics (QED) studies the electromagnetic force on the basis of quantum field theory, so does QCD for the strong force. So, we will find many similarities in applying fields, waves, and interactions, and also in the way the force comes about. However, many of the processes in QCD cannot be calculated exactly. So, our understanding of it is not as advanced as our understanding of QED.

There has nevertheless been a lot of theoretical work and this has been used to explain the many data that have been found to support the ideas of QCD. All in all, it has become a field that provides a well-tested picture of quarks and gluons. This can therefore be taken as a true story of what goes on inside protons and neutrons, as well as the way these particles get together in nuclei.

18.1 The Big Why

Why do we need QCD? What tells us that we have to introduce an extra symmetry (force)?

It all began with the discovery of the particle zoo. One of the denizens was the resonance called the Delta particle. This resonance exists in four forms: Δ^- , Δ^0 , Δ^+ , and Δ^{++} . The idea of postulating quarks came from the search for a deeper mechanism to explain the existence of such particles. It became clear at some point that one could “make” all mesons and baryons in the particle zoo from two or three quarks, respectively, with electric charges $-1/3$ or $+2/3$. For instance, in the case of the Δ -resonance, it can be made from three up and down quarks with the following charges:

- Δ^- = three down quarks (ddd) = $-1/3 - 1/3 - 1/3 = -1$
- Δ^0 = two down quarks and an up quark (ddu) = $-1/3 - 1/3 + 2/3 = 0$
- Δ^+ = one down quark and two up quarks (duu) = $-1/3 + 2/3 + 2/3 = +1$

- $\Delta^{++} = \text{three up quarks (uuu)} = +2/3 + 2/3 + 2/3 = +2$.

Experimentally, they could be made from an interaction between pions $\pi (+, 0, -)$ and protons or neutrons. Pions are made of up quarks or down quarks and their anti-versions. Protons and neutrons are made of up and down quarks. So, a π^+ ($u\bar{d}$) and a proton (uud) could interact by annihilating the d against the \bar{d} . The result is an uuu baryon, hence a Δ^{++} . This works for other baryons and mesons as well, generating a consistent picture that explains how hadrons are made from quarks.

Now we have the obvious question: what holds the quarks together? There must be some force between them. So, we should be looking for a symmetry that generates gauge waves, just as phase shifts generate electromagnetic gauge waves (photons) and SU(2) face changes generate weak gauge waves (W/Z).

There is one hint that we can use to find out what to look for. Experimentally, mesons behave like bosons and baryons behave like fermions. Clearly, quarks must behave like fermions in order to be able to create particle-like behaviour when they clump together into, e.g., baryons. Remember that fermions must make anti-symmetric wave functions and bosons must make symmetric wave functions, so when two fermions are exchanged we find them together in opposite states, and when two bosons are exchanged we find them together in the same state (see Sect. 13.2).

We immediately run into trouble with this idea. If quarks must be fermions and the Δ^{++} particle exists, how can three up quarks be in the same state? Just as there can be only two electrons in the same state (spin up and spin down), two quarks can be in the same state, but not three. Basically, three up quarks must be in a symmetric state, since we get the same state when two up quarks are exchanged. The same goes for the other combinations:

- $\Delta^{++} = uuu = uuu$ (any pair of u-u exchanged)
- $\Delta^+ = uud + udu + duu = uud + duu + udu = \text{etc.}$ (any pair of u-d exchanged)
- $\Delta^\circ = ddu + dud + udd = dud + ddu + udd = \text{etc.}$ (any pair of u-d exchanged)
- $\Delta^- = ddd = ddd$ (any pair of d-d exchanged).

Note that SU(2) allows us to exchange up quarks for down quarks (after all they are two faces of the same particle)! Having said that, a given particle can be made up from any indistinguishable combination of u's and d's. For example, we must make Δ^+ from the sum $uud + udu + duu$, since these three combinations are indistinguishable. What counts is that there is one down quark and two up, but not which one is the down quark (because that is indistinguishable). This sum happens to be symmetric when the u's are exchanged or the d and either of the u's are exchanged. Consequently, all Δ particles seem to be symmetric!

The proton and neutron are different arrangements. They turn out to be of *mixed* symmetry [Ref. 18]:

- Proton = $(ud - du)u$
- Neutron = $(ud - du)d$.

They are symmetric in the sense that $(ud - du)$ can be exchanged with u and we get the same combination ($[ud - du]u = u[ud - du]$). But they are anti-symmetric in the sense that $ud - du = -(du - ud)$. Hence, they are said to be of mixed symmetry.

Mixed symmetry or not, we have to find an additional way to distinguish quarks, otherwise we cannot explain why baryons made from quarks can be fermions. So, we have to find a symmetry that gives us an *anti-symmetric* combination when three quarks are combined. SU(2) does not do the trick, as we have already seen by exchanging u and d.

As it turns out, SU(3) is a symmetry that gives exactly one anti-symmetric combination. In SU(3) we have not two faces, but three faces to turn to. Let's name these three faces after colours: r (red), g (green), and b (blue). We can make an anti-symmetric combination from them:

$$[\text{rgb} - \text{rbg}] = -[\text{rbg} - \text{rgb}] \text{ (b and g exchanged)}$$

But we must add up all indistinguishable versions for a particle. For instance $[\text{brg} - \text{bgr}]$ is indistinguishable from $[\text{rgb} - \text{rbg}]$ since it takes *two* exchanges (exchange b and g and then r and b) to go from $[\text{rgb} - \text{rbg}]$ to $[\text{bgr} - \text{bgr}]$ or vice versa. And two exchanges always lead to the same wave function, whether it is symmetric or anti-symmetric. So, when we add up all indistinguishable combinations we get:

$$[\text{rgb} - \text{rbg}] + [\text{brg} - \text{bgr}] + [\text{gbr} - \text{grb}]$$

This sum is still anti-symmetric. Just exchange any colour (e.g., r) for any other colour (e.g., b) and the sum is the same, but with a minus in front.

Now let's make a product of the symmetries we know: $\text{U}(1) \times \text{SU}(2) \times \text{SU}(3)$. Refer to Fig. 16.2 if you need to recall how to make a product of symmetries. For baryons made from quarks, we saw that $\text{U}(1) \times \text{SU}(2)$ is (partly) symmetric. *But when we make a product with $\text{SU}(3)$ we are able to make a total anti-symmetric wave function out of any combination of three quarks.* In order to make that product, beside a charge ($\text{U}(1)$) and weak isospin ($\text{SU}(2)$), we also have to give quarks a colour charge ($\text{SU}(3)$). These three types of charge together distinguish quarks from each other.

To make this a bit more visible we can try it out for the Δ^{++} particle. The $\text{U}(1) \times \text{SU}(2)$ part of Δ^{++} looked like “uuu”. When we make a product with the anti-symmetric combination in $\text{SU}(3)$ we get:

$$\begin{aligned} \Delta^{++} = & [\text{Ur}, \text{Ug}, \text{Ub} - \text{Ur}, \text{Ub}, \text{Ug}] + [\text{Ub}, \text{Ur}, \text{Ug} - \text{Ub}, \text{Ug}, \text{Ur}] \\ & + [\text{Ug}, \text{Ub}, \text{Ur} - \text{Ug}, \text{Ur}, \text{Ub}] \end{aligned}$$

This is a fully anti-symmetric description of the Δ^{++} particle. Similarly, for other baryons we can create anti-symmetric descriptions by taking the product $\text{U}(1) \times \text{SU}(2) \times \text{SU}(3)$. For protons and neutrons this requires that we first turn the partly symmetric wave function into a completely symmetric one. This is done by adding the spin of the quarks into the wave function. Then this symmetric function is made fully anti-symmetric by making the product with the anti-symmetric colour combination of $\text{SU}(3)$.

Table 18.1 Examples of symmetric and anti-symmetric quark combinations

Quark combination	$u\bar{d}$	$d\bar{d}$	$u\bar{u}$	$d\bar{u}$
Anti-symmetric and spin 0	π^+	η'	H	π^-
Symmetric and spin 1	ρ^+	Ω	Φ	ρ^-

Mesons are bosons. So, they must have spin 0, 1, 2, ... Since quarks have spin $1/2$, mesons must be made from two quarks. As we will see, QCD requires that two quarks can only combine into a free particle when they are a combination of a colour and its anti-colour. Hence, they must be quark and anti-quark. Consequently, colour and flavour team up in the (anti)symmetric combinations we can make. Put differently, the colour does not make a distinguishing factor when we combine two quarks into mesons. Hence, we can restrict ourselves to the $SU(2)$ symmetry to find out whether they are symmetric or anti-symmetric. For a pion (π^+), we have two combinations [Ref. 18]:

- Symmetric: $u\bar{d} + \bar{d}u$
- Anti-symmetric: $u\bar{d} - \bar{d}u$.

But we still have to include the spin. The pion exists only in spin 0. Therefore, the quarks must be spin up and spin down. The anti-symmetric combination can be made symmetric by applying this spin. So, the π^+ is made of [anti-symmetric combination $u\bar{d} - \bar{d}u$] \times [anti-symmetric combination spin up and spin down] = symmetric. So, what about the symmetric combination $u\bar{d} + \bar{d}u$? This exists as well, but combined with the two spins in the same direction (also symmetric). This meson is called the ρ^+ meson and has total spin 1. Other examples are listed in Table 18.1.

Although the whole picture is slightly simplified in order to keep it understandable, the recipe does work for all hadrons that have been measured. Hence, we conclude that we need an $SU(3)$ symmetry in order to describe hadrons as being made from quarks.

18.2 The Colour Symmetry

In the case of $SU(2)$ symmetry, there were two faces a particle could turn into. There were three ways a particle could “rotate”: from one face to the other or back, and a rotation into itself (effectively not changing its face). Before we can continue, we need to look a little more closely at the latter: what does it mean to rotate into yourself?

This is not the same as “staying as you are” since that would not involve a change. We compared the three rotations to the three rotations of a shoe in water. Rotating into yourself would be like the shoe turning through 360° . But the tricky part is that it needs a dimension to make that turn. The three rotations in three real dimensions were just a comparison we made, but let’s go back to the actual $SU(2)$ situation: the only way for a particle to rotate into itself is by turning into the other face and

back again. This only says that we need the dimension of the other face to make the self-rotation. It does not say that we make two rotations: one to the other face and one back. If that were the case, we could see the self-rotation as a linear combination of the two other rotations. So, it is a single rotation in its own right, but it needs the other dimension.

In total, this gives all possible changes in a symmetry containing two states. All possible changes could be expressed as:

$$\text{Number of rotations} = (\text{number of faces})^2 - 1 = 2^2 - 1 = 3 \text{ for the SU(2) case}$$

Quarks are subject to an SU(3) symmetry. That means that each quark has *three* faces it can turn into. The number of rotations is now $3^2 - 1 = 8$. If we wanted to compare these with rotations around an axis, we would need 8 axes and hence an 8-dimensional space. Obviously, we cannot imagine that. So how could we understand this? Let's go back to faces. SU(3) means that there are three faces to turn into. Let's represent each face by one dimension. Figure 18.1 shows these three dimensions and the ways we can change from one face to the other. Each face is assigned a colour, as is customary in QCD. We have red, green, and blue. So, let's count the possible changes, just as in the SU(2) case:

- Change from red to green
- Change back from green to red
- Change from red to blue
- Change back from blue to red
- Change from green to blue
- Change back from blue to green
- Change from one colour into itself via the dimension of one other colour (e.g., from red to red via blue)
- Change from one colour into itself via the dimension of the second other colour (e.g., from red to red via green).

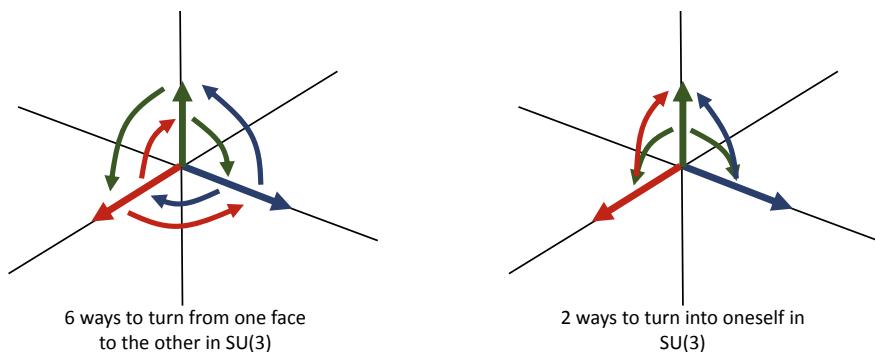


Fig. 18.1 SU(3) represents three states (three faces) quarks can turn into and there are 8 ways to turn. The self-rotations on the right need a dimension to make the change, but do not actually change to, e.g., red and back

These are all the possible rotations we can make when a particle has three faces and there are 8 of them. Consequently, there must be 8 different types of gauge wave associated with each different type of rotation. These gauge waves are called gluons.

The SU(3) symmetry is a global symmetry. This means that the symmetry involves blue quarks and green quarks exchanging colours instantaneously throughout space-time, i.e., with no delays due to the distances between them. As we saw before, the vacuum won't allow changes to propagate faster than c . Hence, the symmetry has to become local and a gauge wave has to carry the change away at maximum speed c . Consequently, all gluons must be gauge waves propagating at the velocity c , unless they are massive. As it turns out, they are not massive. So just like the other forces, we have 8 types of gauge wave, each represented by a type of gluon. They are bosons building up a force field. Each time there is a rotation and a gluon is produced, what happens is similar to what happens with phase changes. The rotation inhibits a potential difference that is propagated away along with the colour change. That potential changes the momentum of the quark that produced it, as well as the momentum of the quark that receives it.

In the case of the W-bosons we saw that W^+ and W^- are each other's anti-particles. That seems logical: to turn from one face to the other is the anti-particle of turning back from the second face to the first. For gluons this goes slightly differently. To understand this, let's see how the changing face works for quarks.

Just as for the weak force, one quark changes face producing a wave that propagates that change away until it finds another quark that changes its face the other way. So, suppose we have a blue quark changing into green. Then the wave must propagate "blue to green" away. More precisely, we say that the blue was taken away and green was created in the quark. So, the wave that propagates the change must carry the blue in it as well as the opposite of green. Let's call the opposite of green "anti-green". Hence, we need an anti-colour for each colour to be able to annihilate that colour. So, the gauge wave must carry blue + anti-green. This way, colour is conserved. This wave can be exchanged only at another green quark (or an anti-blue quark). The gluon uses its anti-green charge to annihilate that quark's green colour and uses its blue charge to paint it blue. So effectively, the blue quark and the green quark will have swapped colours (see Fig. 18.2).

When we look at all the possible combinations of "colour + anti-colour", we get 9 different combinations (three colours \times three anti-colours). But there can only be 8 rotations. So how does that work? Basically, we counted a rotation into itself as three types: blue + anti-blue, red + anti-red, green + anti-green. But there are only two. The essential point here is that a rotation into itself always happens through one other colour and there are only two ways to do that (see Fig. 18.1). For example, turning from blue to blue goes like "blue to red + red to blue" or "blue to green + green to blue". But you might argue that if these are waves, there should be similar waves for "red to green + green to red", etc. However, these waves cannot be distinguished in the effect they create. There are only two ways for a quark to rotate into itself: through the dimension of one other colour or through the dimension of the second other colour. All other ways can be considered linear combinations of those. For

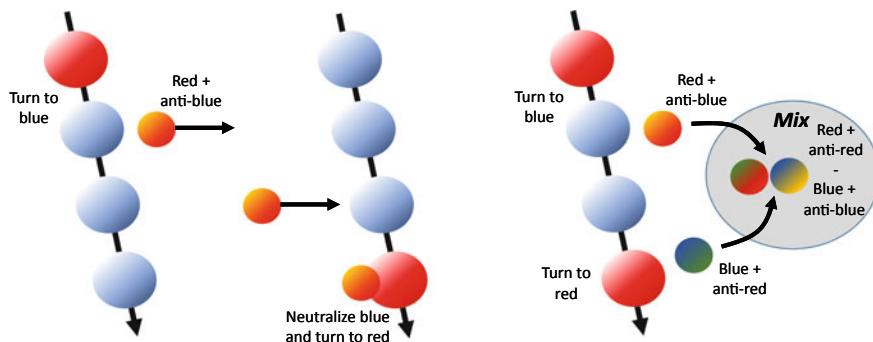


Fig. 18.2 Quarks exchanging colours. On the left: one quark turns blue and the second quark turns red. On the right: a quark turning into itself does not really turn blue, but to make it clear, the change is made from a mix between red + anti-red minus blue + anti-blue. This is a red quark turning into itself “via blue”

example, “red to green + green to red” is equal to a combination of “blue to red + red to blue” and “blue to green + green to blue” (see Fig. 18.3).

In the case of gluons, turning in one direction is the anti-wave of turning in the other direction. Consequently, blue + anti-green is the anti-wave of green + anti-blue. In the section on interactions, we will discuss why Lorentz invariance requires each gluon to be a superposition of a colour–anti-colour combination and its opposite anti-colour–colour combination (e.g., $b\bar{g}$ and $g\bar{b}$). Hence, the 8 gluons are actually made of 6 different superpositions of colour rotations + the two “rotations into itself” [Ref. 74, p. 280]:

1. 6 ways to turn from one colour into the other

- $r\bar{b} + b\bar{r}$
- $r\bar{g} + g\bar{r}$

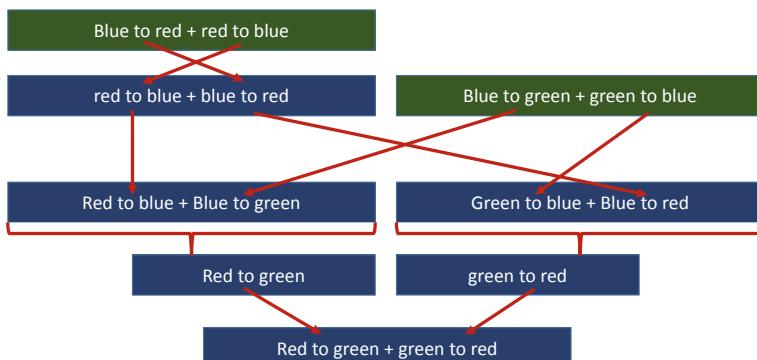


Fig. 18.3 How self-rotations (blue) are linear combinations of only two distinguishable ones (green)

- $b\bar{g} + g\bar{b}$
- $r\bar{b} - b\bar{r}$
- $r\bar{g} - g\bar{r}$
- $b\bar{g} - g\bar{b}$

2. Two ways to turn into itself

There are different possibilities for defining these, but a popular definition is:

- $r\bar{r} - b\bar{b}$
- $r\bar{r} + b\bar{b} - 2g\bar{g}$

By now you may be lost in this colourful world. So, let's try another way of defining the colours and you can pick the insight that suits you best. In this picture, we define the colours using colour isospin I^c_3 and colour hypercharge Y^c . We label the colours in a similar way to the one chosen to label fermions (see Fig. 18.4).

We will see later (see Sect. 18.3) that free particles are colourless. Put differently, all mesons and baryons we make from quarks must have no net colour if they are to exist as free particles. So, the colour of the quarks in these particles must be such that they add up to “white”. In terms of colour hypercharge and isospin, this means that only colour combinations that give $I^c_3 = 0$ and $Y^c = 0$ are possible.

Next, we take the two symmetries of colour and anti-colour and create a product of them. We have done that before, e.g., when we defined $U(1) \times SU(2)$. This basically means that we can make any combination of the two symmetries. We do so in Fig. 18.5.

Making any combination of the three quark colours and three anti-quark colours means that we get all combinations such as $g\bar{r}$, $b\bar{g}$, $r\bar{b}$, etc. The properties of such a combination are additive. For instance, $g\bar{r}$ has the properties of g and \bar{r} added together: $I^c_3 = -\frac{1}{2} - \frac{1}{2} = -1$ and $Y^c = +1/3 - 1/3 = 0$.

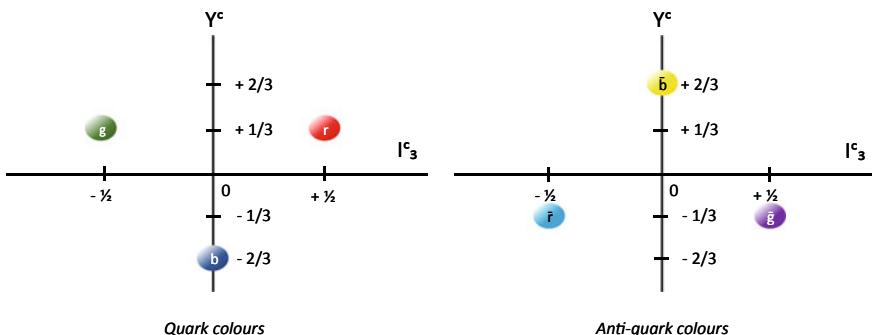


Fig. 18.4 Labelling the colours with colour hypercharge and colour isospin. The colours of anti-quarks are exactly opposite, both in colour and in Y and I . Hence, they are positioned by inversion in the origin compared to the quark colours

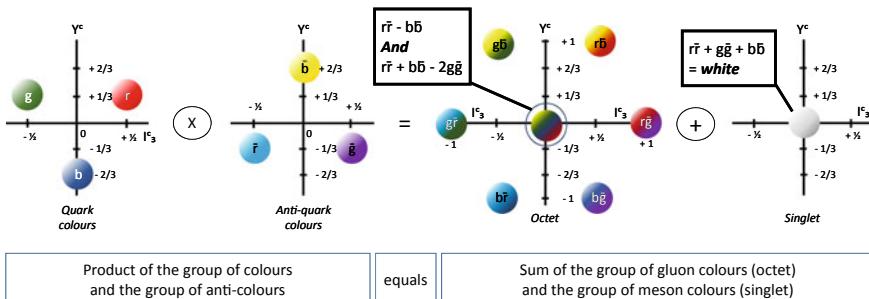
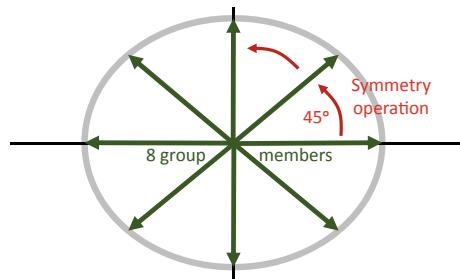


Fig. 18.5 Taking the product of colours and anti-colours means that we can make any combination. But some combinations form a group as we can recombine them into each other (the octet on the right) and some form a smaller group, in this case the singlet on the right

The fun part is that we can find groups for these combinations. One group consists of 8 combinations, which is called an octet (eightfold). Another group consists of just one combination. Why is that? You define a group by means of a symmetry operation. All elements that can be made using the symmetry operation belong to the same group. For instance, when we rotate a vector of length 1 by 45° around some axis, we get another vector that is tilted by 45° compared to the previous vector (see Fig. 18.6). Whenever we apply the same rotation we get a new member of the group, until we carry out the 8th rotation and get back where we started. Basically, by then we have made a 360° turn. So, we can find 8 members of the group that is defined by a rotation through 45° . We cannot reach any other members besides these 8 using the symmetry operation “rotate through 45° ”. So, the group is closed.

It turns out that there is a symmetry operation that links together the combinations described in the octet in Fig. 18.5. All the individual group members are linearly independent, but the symmetry operation relates them. There is one combination that cannot be made from these 8 using that symmetry operation: the combination $r\bar{r} + g\bar{g} + b\bar{b}$. Hence, this combination stands aside from the group. So, we find that there are 9 combinations that can be made from three colours and three anti-colours: 8 of them are linked together in a group and one is not.

Fig. 18.6 Example of a group with 8 group members that are related by a group operation



You may wonder how the two centre objects in the octet can be made from the others. Let's simply count the colours. We could make $r\bar{r} - b\bar{b}$ out of some combination of $r\bar{b}$ and $b\bar{r}$. All the colours we need to make $r\bar{r} - b\bar{b}$ are in these two. We could also make $r\bar{r} + b\bar{b} - 2g\bar{g}$ out of some combination of $g\bar{b}$, $g\bar{r}$, $b\bar{g}$, and $r\bar{g}$. We find two g 's, two \bar{g} 's, one b , one \bar{b} , one r , and one \bar{r} . If we try to make $r\bar{r} + b\bar{b} + g\bar{g}$, we find that we can never get this combination together. There are always surplus colours remaining or some colours missing. So, the combination in the singlet cannot be made from the 6 around the octet using the symmetry operation that makes the octet into a group.

The octet has 8 combinations that have a colour (the six around the centre) or are colourless (the two in the centre). The 6 combinations that have a colour have exactly the properties we expect from gluons. When a quark changes from red to blue, it must propagate away the red and anti-blue. This is one of the combinations in the octet! In the centre we find two white combinations. They look exactly like the two self-rotations we identified earlier. Since they are linked by a symmetry operation to the other combinations, they can describe a rotation into itself via another colour. Hence, the two combinations in the centre can describe self-rotations. So, the whole octet can be used to describe all gluons! That is, it can describe the colour part of the gluon wave function.

What about the singlet? Can this describe a gluon too? Basically, this combination describes "stay as you are". It is not only white, but it can also not be made from the other combinations. So, it does not represent *any* change. When a quark stays as it is, it does not produce any gauge wave. So, this combination describes a trivial situation when it comes to gluons. It does nothing. If it would be a gluon after all, it would be white and would not interact with anything else. So, it cannot be produced (since it does not propagate a rotation) and it cannot create a rotation in any other particle.

So, is the singlet entirely useless? Not in a different context! Let's look at mesons. They consist of a quark and an anti-quark. Quarks always have one of three colours. Anti-quarks always have one of three anti-colours. However, mesons can only exist as a free particle when they are white. So, it would seem as if the octet could describe colour and anti-colour combinations, but these are not white. Mesons cannot be white and turn into non-white by means of a symmetry operation. Hence, the white combinations in the centre of the octet cannot be used in a viable meson wave function either. Mesons must be white at all times. As soon as they become non-white by any means, they will start to interact through the strong force and quickly clump with other particles until they become a particle that is truly white (see confinement). Truly white is exactly what is described by the singlet. So, the singlet is the perfect candidate to be the colour part of a meson wave function.

Remember what we said before about mesons: "colour and flavour team up in the (anti)symmetric combinations we can make" (see Sect. 18.1). And we basically disregarded the colour as a distinguishing factor for the symmetry of the meson wave function. The symmetric wave function we needed for mesons we could make entirely from spin and SU(2) combinations. Now we see the reason: the singlet colour state is the only combination that fits a meson and it describes a symmetric wave

function. And when it comes to symmetric parts of the wave function, they leave the other part of the wave function alone. Hence, symmetric \times symmetric = symmetric and symmetric \times anti-symmetric = anti-symmetric.

Let's look a little deeper into the colour wave function for a meson. It is made from a linear combination of $r\bar{r}$, $b\bar{b}$, and $g\bar{g}$. So, the quarks in the meson can change their colour combination, but will always have a colour + its opposite anti-colour combination at any point in time. We can see this as follows: start with the quarks in the meson having the colour r and \bar{r} . Suppose the r quark changes to b (blue). It produces a gluon made from the combination $r\bar{b}$. This gluon finds the other \bar{r} quark. It annihilates the \bar{r} colour and paints that quark \bar{b} . So now the meson turns from $r\bar{r}$ to $b\bar{b}$. When a meson always has a colour + its exact opposite anti-colour combination, the colour does not play a role in determining the (anti)symmetry of the meson.

Summarizing, we took a combination of a colour and an anti-colour and we found two groups of such combinations that could actually describe the colour part of the wave function for gluons and mesons. That's fun! Let's try some more! How about we make a combination of a colour and a colour. What would that tell us? (see Fig. 18.7).

The combinations we can find we can again group into two groups. Could these groups represent anything? Let's start by remembering that quarks only carry colour (no anti-colour) and only anti-quarks carry anti-colour (no colour). So, the quark combinations we could describe with a colour + colour combination are only quark-quark combinations. Anti-quarks will not be involved.

The first thing we notice is that there is nothing white to be found in these groups. Hence, we cannot describe a free particle using these groups. Consequently, we can say that colour charge does not allow quark-quark combinations to exist!

The second thing we could investigate is whether the colourful combinations could describe any kind of gluon. In theory they could, but the interesting question is: how could we produce such gluons? A gluon is produced when a quark changes colour or when an anti-quark changes anti-colour. In both cases the change can only be propagated away by a colour-anti-colour combination. None of the groups in

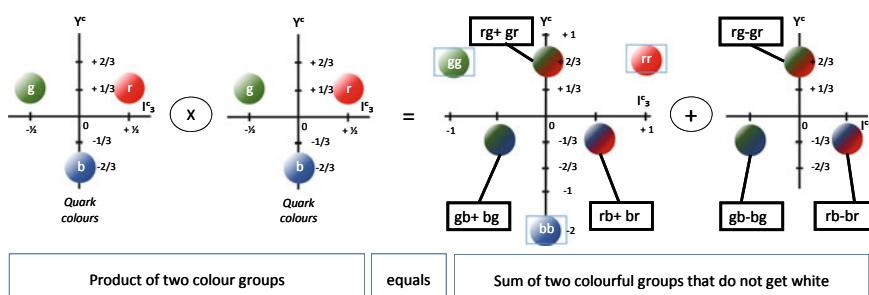


Fig. 18.7 A product of quark colours with quark colours (qq). This leads to all colourful combinations but no colourless ones. So, a quark could not pair up with a quark as it would still have colour and quickly attract gluons and another quark

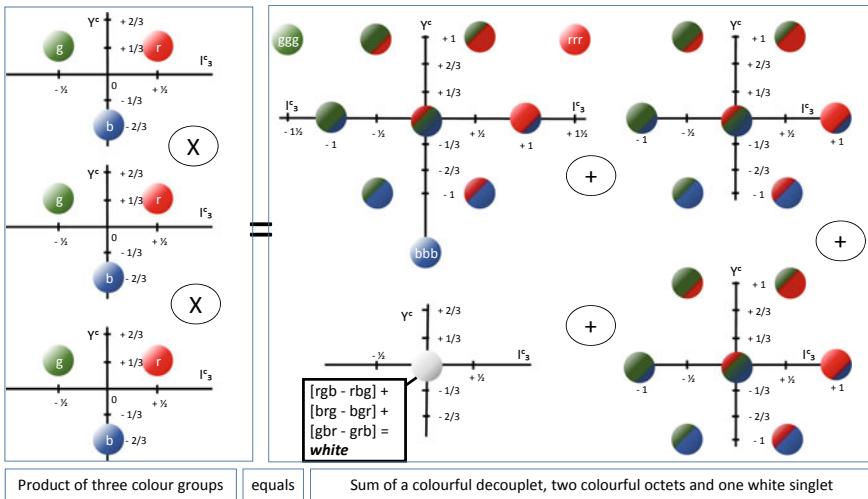


Fig. 18.8 Three quark colours getting together produce a product of “quark colours \times quark colours \times quark colours” (qqq)

Fig. 18.7 describe colour–anti-colour combinations. So, colour–colour combinations may describe a type of gluon, but we cannot produce it! Hence, these do not exist either.

So, let’s raise the stakes and look at combinations of three colours instead of two (see Fig. 18.8). Again, these can only describe quarks. Anti-quarks will not be involved.

There are a lot of combinations one can make from three colours. Some are symmetric, some are partly symmetric, and some are anti-symmetric. We can find four groups of such combinations. Each group has a white combination in its centre. However, the same applies to the white combinations we met before. When there is a symmetry operation that can turn the white combination into a different non-white combination, it is not truly white. Consequently, it cannot describe the colour part of a free particle wave function. Again, only the singlet can describe a truly white colour wave function. It turns out that the combination that belongs to the singlet is the anti-symmetric colour wave function we found before. Hence, this is the only wave function we can use to describe the colour part of a qqq combination of three quarks. Consequently, for such a qqq combination to always be anti-symmetric (as fermions must be), the other parts of the wave function must be symmetric, as we saw before.

We can do the same for three anti-colours. This will give us the single solution to describe anti-quarks. We do not have to look at gluons, for the same argument applies as before: the colourful combinations cannot be produced by any rotation we know of. Any other colour combinations we could make do not add anything as they all give either non-white + non-singlet solutions or colour combinations that cannot be produced and therefore cannot represent viable gluons.

We started all this from the conclusion that we need an SU(3) symmetry in order to describe a viable three-quark fermion (anti-symmetric) wave function. But we could also describe a viable quark–anti-quark boson (symmetric) wave function to represent mesons.

The SU(3) symmetry is non-Abelian, just like SU(2). This means that two rotations one after the other performed in different orders do not generally give the same result (remember the example of the dice we discussed before). Just like the SU(2) case, we can achieve the same end result for two turns in different order. This can be done by propagating away an extra change *within* the gauge wave. Consequently, gluons can create such changes by themselves. This means that gluons can interact with each other (they feel the changes themselves). It leads to an interaction potential between gluons and we will see an interesting consequence of this behaviour later.

Finally, it is important to note that the SU(3) colour symmetry is an exact symmetry. By comparison, SU(2) was an exact symmetry before it got broken by the condensation of Higgs. In today’s universe it is no longer an exact symmetry. We have seen the consequences of that, such as the fact that flavours can mix and as a result can turn into another face than the one expected from the symmetry. This does not happen in SU(3). Hence, the strong interaction does not differ with respect to different colour rotations. Each rotation has the same strength.

18.3 QCD Fields: Overview

The QCD overview of waves contains a wave for each quark and a gauge wave for each gluon. Quarks also connect to U(1), hence they have charge. We saw before that they also connect to SU(2) and so they have isospin as well. These two symmetries lead to many interaction potentials for the quarks that we have basically already discussed. Therefore, we leave their fields out for now and focus on the colour force between quarks (see Table 18.2).

The total number of quarks is $6 \text{ flavours} \times 2 \text{ handedness} \times 3 \text{ colours} = 36$ quark types. They all couple to 8 gluon gauge waves by the same strength. So, the coupling strength is entirely symmetric across all quark types. The same goes for the 36 anti-quarks, so in total we have 72 quark types, all symmetric with respect to colour charge.

Table 18.2 Fermion and boson fields in QCD and their mutual interactions

Field	Wave type	Mass “spring”	Interaction “spring”
6 right and 6 left-handed quarks (u, c, t, d, s, b)	Fermion	Different for each quark flavour	With gluons, photons, W, and Z°
8 gluons	Gauge boson	0	With quarks and other gluons

Gluons can interact with each other. Just like the weak bosons they are able to rotate in the colour space by themselves generating new gauge waves. This is a non-linear behaviour, causing ever more gluons to appear at larger distances. The weak force did not show the consequence of this since its bosons are too heavy to reach any distance. However, gluons do not connect to Higgs. The Higgs field does not feel colour charge, and so when it got broken, this did not result in mixing with gluons. Consequently, gluons remain massless. The effective potential between quarks can be computed as a function of their distance, as a consequence of this non-linear behaviour. Since this effect is large for massless gluons, it leads to behaviour such as “colour confinement” and “asymptotic freedom”.

18.3.1 Colour Confinement

We have seen what happens when we take two quarks and consider the colour interaction between them. A quark changing its colour produces a gluon that changes the colour of another quark (see Fig. 18.2).

The colour change creates a potential difference that is propagated away by the gluon, just like the electromagnetic force. However, there is also an important difference. The potential between two quarks is a complicated matter. It can be attractive or repulsive, depending on the colour combination. For instance, the colour singlet state that describes a meson with one quark in one colour and the other in the same anti-colour yields an attractive potential between the two quarks. But if we carry out this calculation between two coloured quarks that do not make a white colour together, e.g., red and anti-green, the two quarks will experience a repulsive potential. This applies to the octet in Fig. 18.5. This behaviour is in line with the idea that colourful particles cannot exist and hence the octet does not describe possible quark-colour combinations. According to these calculations, quarks will only attract each other when they form a white combination.

Over a longer range, the potential between two colours is determined more by gluons producing other gluons and interacting together. Beyond the proton diameter, the resulting average potential is constant *per unit distance*. Hence, the greater the distance between two quarks, the stronger the potential, and this relation between potential and distance is linear! It gives a potential of about 1 GeV per femtometre. The associated force is equal to lifting 16 tons.

Now suppose we try to pull the two quarks apart (imagine attaching 16 tons to one quark). The effect is that the distance between them is increased. The gluon waves that are exchanged between them have to travel a greater distance. Over that distance there is more chance for a gluon to produce an extra gluon. Since gluons are creating an attractive force between the quarks and between each other, the attraction between the quarks will have gone up. The further we manage to pull them apart, the more time a gluon has to produce other gluons, and more gluons means a stronger attraction between the quarks.

This process can continue since the gluons are massless. Therefore, it does not cost much energy to produce new ones, in contrast to the situation with the weak force, where the massive bosons cannot be produced in large quantities. So, we can imagine how different the “weak” force would have been before symmetry breaking, when the W-bosons were still massless!

As we continue to pull the quarks apart, there are ever more gluons between them. Since the force increases, we need to put more and more energy in to pull them further apart. That energy gets stored in the ever-increasing numbers of gluons, until there is enough energy stored in them to create a new quark–anti-quark pair. When that happens, the pair of quarks that attracted each other before will now have become two pairs of quarks that attract each other. Hence, a quark can never be pulled out of a bound state of quarks. Pulling quarks apart always costs more energy than needed to create new quarks that bind with the quarks that are pulled apart.

Suppose there were a loose quark of some colour. It would produce gluons and those gluons would multiply like rabbits the longer they have to fly about out there in search of another quark. So, any other loose quark will be found and pulled to the loose quark with great force. Once again, we conclude that it will be impossible to find a loose quark.

Quarks will also be synchronized in colour. Suppose we have a red quark and a green quark together. Suppose further that they produce gluons that bind the two quarks together. However, together they present a net colour and so they will still produce gluons that fly out in search of other quarks. Only when no net colour charge is presented to the outside world will there be no gluons flying out in search of other quarks.

Consequently, *any which way we look, we find quarks tightly bound together in combinations that have no net colour charge. This is called confinement.* There are two types of combination that have no net colour charge and we met them before: mesons (quark + anti-quark) and baryons (three quarks with colours r, g, and b).

Quarks do not actually have a colour. It is just that the parallel between the way primary colours are added and the way the three different types of charge add up are so similar that colour serves as a nice analogy.

There is one more possibility to form a colourless combination. This does not require quarks, but only gluons. After all, since gluons feel the same force, they can group together as well. Such combinations are appropriately called “glueballs”. It is hard to verify their existence experimentally, but they could exist.

This idea also tells us how gluons behave between quarks. We find most gluons between two quarks. Consequently, the other gluons produced by these gluons will feel a net attractive force towards the quarks and the gluons in between them. Soon there will be many gluons between the quarks. Therefore, the strong force between quarks tends to clump together in gluons along the line between two quarks. Hence, the interaction between the quarks inside these particles can be compared with an elastic band that pulls more strongly the more it gets stretched (see Fig. 18.9). Such an elastic band is called a “flux tube” or “string”.

Compare this behaviour to the electromagnetic force. When we move further away from a charge, the gauge photons produced by it will diminish in number

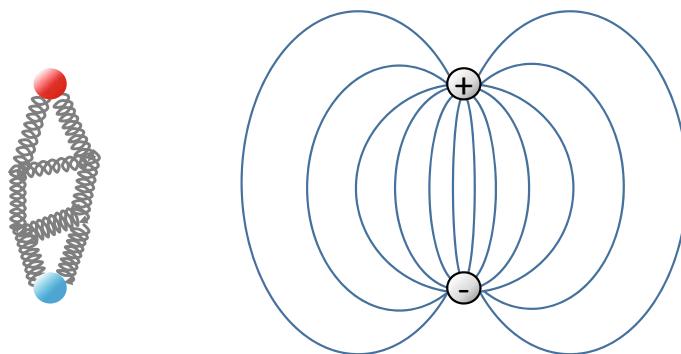


Fig. 18.9 The behaviour of gluons that attract each other leads to a grouping of gluons between two quarks (left). Consequently, they behave like a rubber band that pulls back harder when stretched. Photons do not attract each other and hence can spread through space, so electromagnetic effects become weaker at long range, shown here by the way field lines get further apart at greater distances (right)

since they get spread over a larger area. Hence, the field will decrease with distance and consequently, the effective force will also decrease with distance. In contrast, the “elastic band” behaviour of the gluons means that they do not spread out with distance. Hence, the force will remain the same when the distance between quarks grows. There are only more and more gluons on the way, containing more energy in the band, just as an elastic band that gets stretched contains energy.

However, an elastic band could also snap. This is what happens when the energy content is large enough to produce a new quark–anti-quark pair, as discussed above. A meson that gets stretched this way soon consists of two mesons. So, the band snaps by creating new particles.

So, what happens when two or three quarks are living happily together in a composite particle? All gluons produced in any other direction than the line between two quarks are being pulled towards the quarks as well as the line(s) between the quarks. So, the gluons tend to accumulate on these lines. Consequently, we do not find any gluons outside the particle. Or rather, the strong force will diminish very quickly outside the particle.

In conclusion, we can say that the effect of gluons attracting other gluons leads to a strong force between two or three quarks as well as the rapid reduction of force outside this group of quarks. This effect is called colour confinement. It means that quarks cannot be found as loose particles, but it also means that gluons do not appear outside the particle. So, all colour from quarks and gluons is assembled together in a particle that shows no net colour to the outside world. This effect limits the range of the strong force to the order of 10^{-15} m, which is about the size of a nucleus. The strong force plays a role in pulling together protons and neutrons in a nucleus, but we will discuss that in the sections on composite particles and interactions.

18.3.2 Quark Jets

Suppose you shoot a high energy electron at a proton. You give the electron so much energy that it can enter the proton and fly “between the quarks”. An electron does not feel the strong force, so it does not interact in that way. You only need to overcome the electromagnetic force and the weak force (!) So sometimes the electron gets scattered by the charge of the quarks and sometimes it meets a W- or Z-boson and interacts via the weak force. But sometimes it may succeed in transferring a large amount of momentum to one of the quarks. What happens then? [Ref. 22].

The quark gets shot out of the proton. When a quark accelerates, it starts to emit gluons. Just as an electron (or any charged particle) that gets accelerated starts to radiate photons, a quark will do so as well. But the quark will not only radiate photons, it will also emit bosons from the other forces it couples to. So, gluons and W/Z-bosons are radiated as well. If the accelerated quark is a top quark, it will emit a W-boson (and turn into a bottom quark) before anything else happens. But for the other quarks, the W/Z are too heavy and they radiate many gluons.

These gluons fly off at high velocity. The velocity of the gluons is in the direction of the kick that the quark received from the electron earlier. The gluons feel the strong force just like the quark, so between all the gluons flying off, bands of other gluons are formed. This process goes very fast. At some point the various bands of gluons break. When they break, they form new quark–anti-quark pairs out of the energy stored in those bands (see Fig. 18.10). Consequently, we see many mesons fly off around the direction of the original gluons. Most of the energy that is produced in these processes goes into the velocity (kinetic energy) of the particles that are produced.

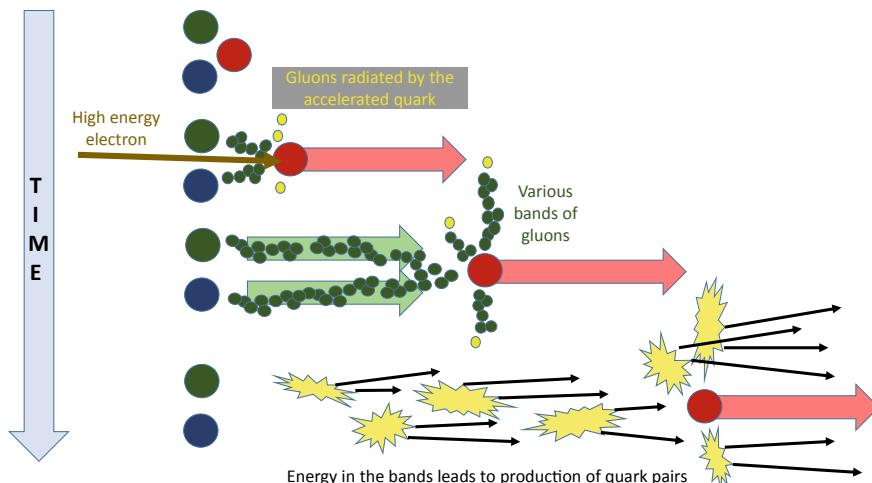


Fig. 18.10 An ejected quark radiates gluons. Between all the gluons and the quark, bands of new gluons form. When they break, they form jets of mesons

Since momentum and energy are conserved, the total energy and direction of motion of the jet particles sum up to the total energy and momentum of the quark that was kicked out. So, by measuring the particle jet we can estimate the velocity and direction of the quark that got ejected. Charge is conserved as well (of course), so the particles that are produced in the jets must be electrically neutral. Hence, they can be quark-anti-quark pairs.

Particularly interesting are three-jet events. In these events mesons are broken apart with two quarks flying off in different directions. This can for instance happen as a part of a jet event. Sometimes three jets of hadrons come flying out of such an event. How can we produce three hadron jets? This can only happen when there are three particles flying off that each feel the strong force. So, we would have three bands form between the three particles, and each would break and produce a jet of hadrons. The problem is that it is primarily quark-anti-quark pairs that are produced in these jet events. So, when the quarks in such pairs have enough energy to move apart, they can produce one or two jets, but not three (see Fig. 18.11). For a third jet to form, three particles must fly off instead of two. Two of those are the quarks that break apart. So, what is the mysterious third particle?

That third particle can be viewed as a gluon or a photon. Between the quarks, but also between the gluon and each of the quarks, a new band of gluons is formed until it breaks. Then hadrons come flying out in three directions (see Fig. 18.11). If the third particle was a photon, only one jet would appear as a result of the band breaking

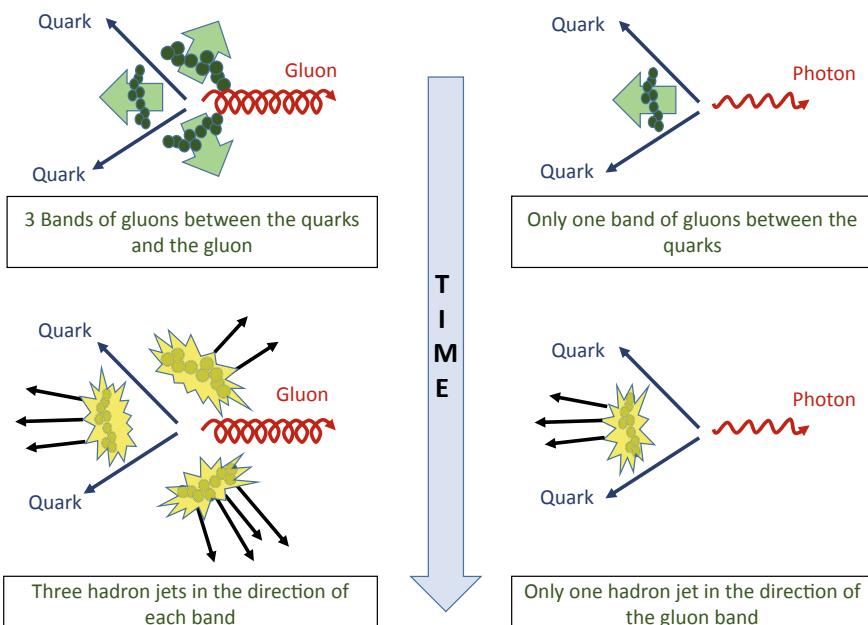


Fig. 18.11 Three-jet events with a gluon (left), compared to what happens when the third particle is a photon (right)

between the quarks. Experiments show that three jets come flying out, proving that a gluon was among the original three particles.

Three-jet events were observed and extensively studied first at the DESY laboratory using the PETRA accelerator [Ref. 77] and later also at CERN using the LEP (Large Electron Positron collider) [Ref. 78]. These observations are the best proof we have for the existence of gluons.

18.3.3 Asymptotic Freedom

When we discussed renormalization, we saw that the electric charge we measure goes up when we close in on a charged particle. As the length scale gets smaller, the charge tends to infinity.

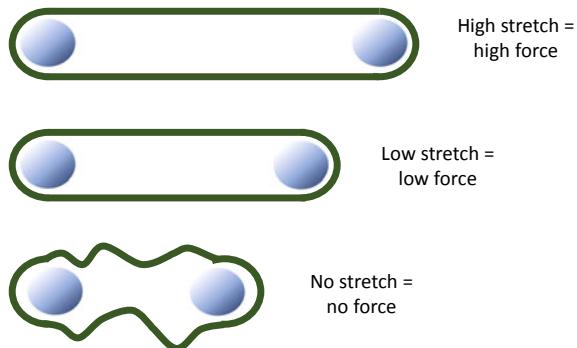
It was also explained that QCD is essentially different: the colour charge does not go to infinity when we move in to a zero length scale. So, let's see how that works. Gluons can produce a virtual quark–anti-quark pair. Just as virtual photons do in the presence of an electric charge, we may expect these virtual quark–anti-quark pairs to screen off the colour charge of the bare quark. However, the colour charge is different in a number of ways.

First of all, a gluon changes the colour of a quark. What does that mean in terms of polarization? In electromagnetism, positive and negative charges are attracted and repelled, respectively, by a negative charge. Hence, polarization can occur. However, in colour charge, all coloured particles (quarks and gluons) attract each other while exchanging colours. So, a quark–anti-quark pair does not feel a different force from one colour or another, unlike an electron–positron pair, which would experience repulsion resp. attraction in the vicinity of an electron. In the attraction process, colours get exchanged, so quarks change colour all the time, again unlike electric charges, which do not change when they interact. The result is that the screening effect is not relevant for quarks! The gluons swarming around a quark will only change the colour of other quarks and attract them in the process. They will not polarize colours (e.g., blue in one corner and red in the other) as electric charges would.

Second, we have seen how gluons produce more gluons at greater distances. So, you can imagine what happens at shorter distances: there will be fewer gluons. In fact, the closer we get to the bare quark, the weaker the force gets. Typically, the strong force remains constant (per unit length!) at distances greater than the size of a proton. But it decreases in strength below that size. This makes sense, as the size of a proton would be determined by the strength of the colour interaction. This effect is opposite to the screening effect we saw in the case of electrically charged particles. Hence, it is sometimes called an anti-screening effect. According to this effect, closing in on a quark decreases the force because fewer and fewer gluons appear at such distances.

At higher energies we get closer to the bare quark. Suppose the vacuum has a high overall temperature, comparable to such a high energy. Everything that swarms

Fig. 18.12 An elastic band between two quarks. At small distance, the quarks are free because the band imposes no force



around in the vacuum has on average that very high energy. Then all (gauge) waves would be able to get really close to a quark. The quark does not feel colour charge so much at that distance. We can compare the gluon behaviour with an elastic band between two quarks, getting stronger when it is stretched and weaker when we get closer (see Fig. 18.12) [Ref. 22].

Hence, the environment experiences quarks as if they were free. The environment interacts with quarks in a way that does not get overwhelmed by the strong force. The temperature at which this would happen is of the order of 2×10^{12} K, or on average an energy of 130–140 MeV per particle. This energy level is called the Hagedorn temperature, after a German physicist named Rolf Hagedorn who discovered this at CERN in the 1960s [Ref. 79]. Such energies can easily be reached by modern particle accelerators. Consequently, what happens at this temperature has been extensively researched.

At such temperatures quarks and gluons become “free” and form a kind of quark–gluon plasma, sometimes also called “quark matter”. In this state, the quarks do not group into hadrons, so this might be viewed as a “boiling point” of hadrons above which hadrons can no longer exist. This state is reached in particle accelerators, but only for a short time. The energy quickly spreads and the temperature drops. After that the quarks get confined again into hadrons. Consequently, many new hadrons are formed that quickly move away from the collision point.

The freedom that quarks experience at such high energies is referred to as “asymptotic freedom”. It means that when the distance approaches 0 (the zero-distance asymptote), the quarks and gluons become free. Clearly, there is no infinity problem in QCD with this behaviour. That means that there is no objection against any small length scale in QCD and the theory may be correct even way below the Planck scale. However, that does not mean that it is. If the vacuum has a structure at, e.g., the Planck scale, QCD will not work below that scale. After all, QCD is a quantum field theory that describes quarks and gluons as waves and disturbances and these will not exist below the scale at which we start to see the structure of the vacuum itself.

18.4 Composite Particles

We have learned a lot about quarks and gluons. We know they make up hadrons (mesons and baryons). So, let's take a look at the types of particles that can be made by combining quarks. We start by considering particles that are made from just three quark flavours: up, down, and strange. We choose these as they lead to the most commonly produced particles in accelerators. They are also the particles first discovered in the “particle zoo”.

When we combine three flavours, we can treat such combinations in a similar way as we did when we created groups from three colours. The first group we look at are the mesons. Since they consist of a quark and an anti-quark, we use the same type of group product as that of colour \times anti-colour. However, we are using different quantum numbers. We will use the charge of the quarks and a number S, which is called “strangeness”. That number is in fact nothing other than the flavour of the s quark. So, this number only says something about the total number of strange quarks in the combination. The reason why we use Q and S is that we can then distinguish between the three flavours u, d, and s that we want to find combinations of. Note that the down and strange quarks have the same charge, weak isospin, and hypercharge. The only way to distinguish them is by means of their flavour. The result is shown in Fig. 18.13.

The result is grouped differently, this time into a nonet, so it is not split into an octet and a singlet. The reason for that is that *all* 9 possible combinations of three flavours are related by a symmetry operation. That is, we can change the up, down, and strange faces of quarks into each other. Consequently, the three flavours group together according to a different group than SU(3). This group is called U(3), but we will not go into the details here.

In the general literature the Q- and S-axes make an angle such that the two balls on the left shift down and the two balls on the right shift up. This is done to make the nonet look like one of the octets we saw before. This suggests more of a symmetry.

The total wave function of mesons is made of a spin part, a flavour part, and a colour part:

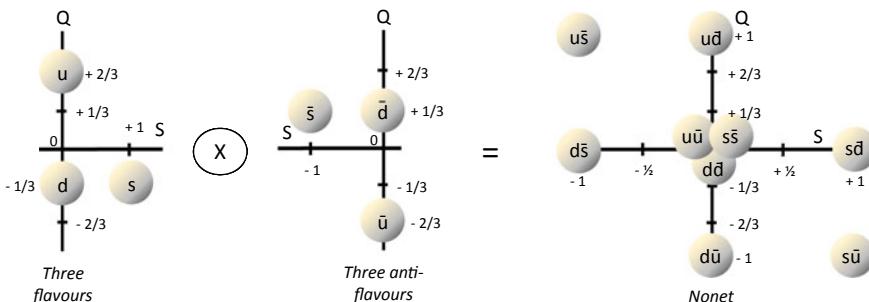


Fig. 18.13 What happens when we combine two groups of three quark flavours: we get a nonet

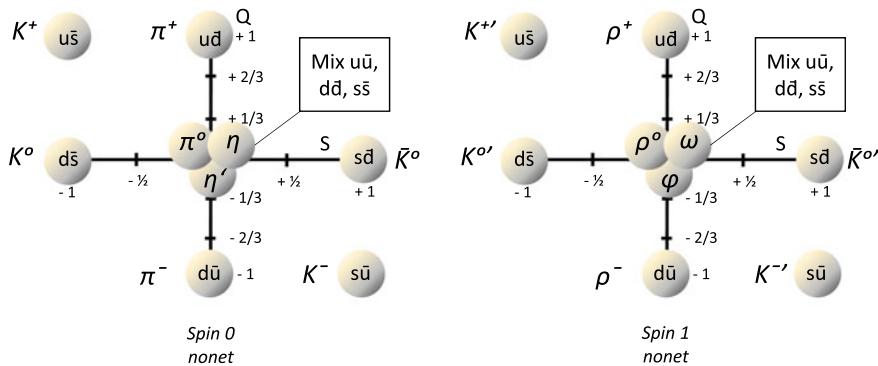


Fig. 18.14 The two meson nonets for spin 0 and spin 1, for mesons made from u, d, and s

$$\begin{aligned} \text{Meson wave function} = & (\text{spin } 0 \text{ or } 1) \times (\text{2-flavour combination}) \\ & \times (\text{meson colour combination}) \end{aligned}$$

We saw before that the colour part is only one aspect of the wave function. Spin adds two types. Each (anti-)quark in a meson carries a spin $\pm\frac{1}{2}$. So, we can have spin 0 ($\frac{1}{2} - \frac{1}{2}$) or spin 1 ($\frac{1}{2} + \frac{1}{2}$). Consequently, when we look at the flavour nonet, there are actually two such nonets to represent mesons made from u, d, and s. One nonet with spin 0 and one nonet with spin 1 (see also the table of mesons in Sect. 18.1). The two nonets are shown in Fig. 18.14.

We can also look at the group of baryons. In this case we group three quarks together. That means we will look at the product group of three flavours \times three flavours \times three flavours. This looks like the grouping we did for colours in Fig. 18.8. For baryons, the total wave function is made of three parts:

$$\begin{aligned} \text{Baryon wave function} = & (\text{spin } \frac{1}{2} \text{ or } \frac{3}{2}) \times (\text{3-flavour combination}) \\ & \times (\text{baryon colour combination}) \end{aligned}$$

With three quarks, we can have spin $\frac{1}{2}$ and spin $\frac{3}{2}$. This leads to two groups: a decuplet for spin $\frac{3}{2}$ and an octet for spin $\frac{1}{2}$. The difference is that the combinations uuu, ddd, and sss all have spin $\frac{3}{2}$, but cannot exist in a spin $\frac{1}{2}$ mode as this would lead to a symmetric wave function (including SU(3)). Therefore, we find these in the decuplet and not in the octet. The result is shown in Fig. 18.15.

The octet contains two types of combination “uds”. The Σ has isospin 1, and the Λ has isospin 0. Otherwise, they are the same.

We can extend this exercise to all the quark flavours. When we extend to four flavours, we get three-dimensional models (so-called super multiplets) of flavour combinations. You can imagine that with six flavours we end up with many more particles than we discussed in the particle zoo. To differentiate them we need quantum numbers for each flavour. In this way we can create all baryons and mesons from

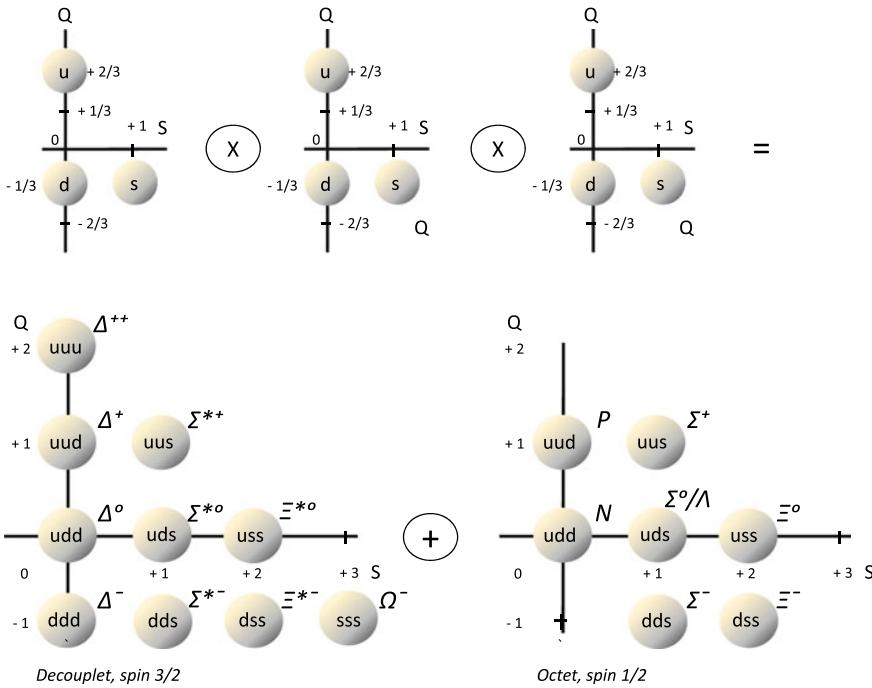


Fig. 18.15 The two baryon groups originating from the product of the flavours u, d, and s

six quark flavours and their spin. Their charge and weak isospin are just found by addition of the charge and isospin of the individual quarks they are composed of.

18.5 Interactions

Let's look at the different ways quarks and gluons can interact. This is more complicated than, e.g., photon interactions, since we have 8 types of gluons and we have three colours and three anti-colours to take into account. In the regular Feynman diagrams, a gluon is depicted as a “bulb-wire” instead of a wavy line (like photons and W/Z bosons). However, in order to keep track of the colours and gluon types, a more interesting way to depict them is by drawing a line for each colour. A quark has one colour, hence one line. A gluon has two colours and therefore two lines. The colour has an arrow to the right and an anti-colour has the same colour but an arrow to the left. With these rules we can keep track of the colours and which gluons we are dealing with. In the diagrams below we have used both methods so that you can see the difference.

18.5.1 Quark–Quark Colour Interactions

Quarks bind in mesons and baryons by means of the colour force. This means that the potential they feel from the other quarks must be propagated by gluons. They do that by producing a virtual gluon, much as an electron produces a virtual photon. So, the quark wave splits into a quark disturbance and a gluon disturbance. The split leads to a virtual gluon carrying a certain colour combination, while the virtual quark has a potentially different colour. Since a gluon is massless it does not necessarily require a lot of energy to produce the gluon disturbance. The gluon disturbance gets either reabsorbed, resulting in the same quark, or the gluon gets absorbed by another quark. In the latter case, the original quark becomes a real quark wave but with a different energy and momentum when the gluon is absorbed by another quark. Consequently, we can say that the energy for creating the gluon was taken out of the momentum of the original quark. Hence, quarks emitting gluons lose momentum. This concept is not very different from the electron case.

Let's first take the case of a quark changing colour and producing a gluon that is absorbed by another quark. Basically, the effect of this interaction is that the quarks will have swapped colour (see Fig. 18.16).

We saw before that a gluon is a sum of a colour + anti-colour *and* vice versa. In the example of Fig. 18.16, the gluon is a combination of red + anti-green ($r\bar{g}$) *and* green + anti-red ($g\bar{r}$). We said then that $r\bar{g}$ cannot be distinguished from $g\bar{r}$. When we look at the colour streams in Fig. 18.16 we can see why. In the diagram there are two vertices: one where the gluon is produced and one where it is absorbed by the other quark. These two vertices are events that can be “observed” in a different order by different observers. The observer of the colour stream in the top right diagram concludes that the gluon must be $r\bar{g}$. The observer of the colour stream in the bottom right diagram concludes that the gluon must be $g\bar{r}$. So basically, $r\bar{g}$ is the anti-gluon of $g\bar{r}$. When the order of events is different to another observer, the interaction must still be the same. Put another way, the interaction must be invariant

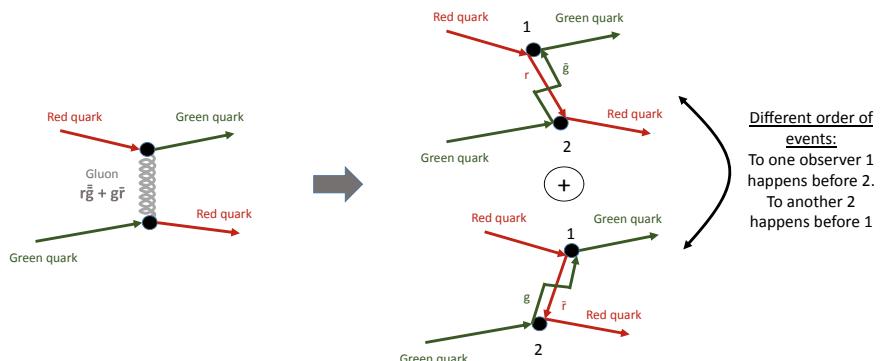
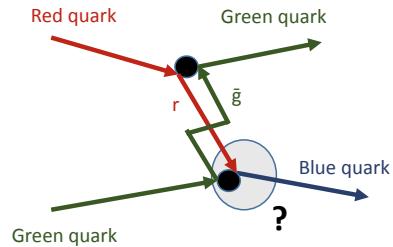


Fig. 18.16 Two quarks swapping colour by exchanging a gluon

Fig. 18.17 An impossible diagram, illustrating that colour must be conserved



under a Lorentz transformation. This means that, when observed by someone with a different velocity, the interaction must be the same. Hence, the interaction can only be invariant under Lorentz transformations if we describe it as a combination of both $r\bar{g}$ and $g\bar{r}$. Consequently, both observers are included in the description of the gluon.

When we observe the colour lines, we also see that each colour follows a line that is not broken throughout the interaction. This indicates that colour is conserved in the process. In all the diagrams we find that colour lines pass through unbroken. To illustrate this, see Fig. 18.17. In this diagram, colour is not conserved in this sense. At the point indicated by the question mark, a colour line is broken. But there is no process that could create this situation, so this diagram is not possible.

The next situation to explore is when quarks do not change colour. This is the “rotation into itself” type of interaction. Figure 18.18 shows three examples of interactions that leave the colours of quarks unchanged. In the leftmost diagram a red quark can interact with a red quark while keeping the same colour by means of two types of gluon. These are exactly the two types of gluon that can be produced by a self-rotation. The colour stream illustrates what part of the gluon is necessary for the interaction. Both types have a $r\bar{r}$ part in them so they are both applicable. You might wonder what happens, e.g., to the $b\bar{b}$ part of the gluon. This part can be modelled as a loop. So, the blue is not produced by either quark and not absorbed by either quark. This is a different way of illustrating how the quark that produces the self-rotation needs the blue dimension, while the quarks do not really turn blue in the process.

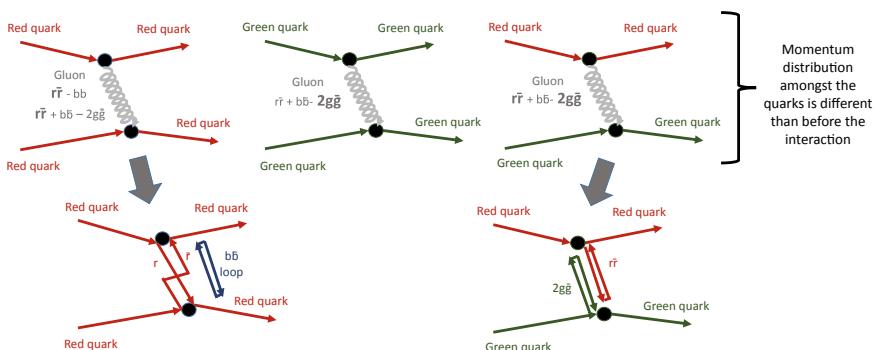


Fig. 18.18 Several self-rotation diagrams

The middle diagram illustrates two green quarks exchanging a gluon. In this case, only one type of gluon can be exchanged. This is the type that carries the “ $-2g\bar{g}$ ” in it. The other type does not carry green, so it cannot facilitate this process. You may wonder whether this means that two green quarks would have a lower interaction rate compared to two red quarks. After all, the interaction strength grows when there are more types of gluon available to carry the transaction. The point is that the “2” in front of the $g\bar{g}$ in the combination implies that the green–anti-green colour line is of “double strength”. So, there may be only one type of gluon for a green interaction, but it is twice as strong. As a result, all colours interact equally strongly.

The diagram on the right illustrates that self-rotations can also interact with quarks of a different colour. The two colours remain the same, as can be seen in the colour stream diagram. In all cases the interaction does transfer momentum and energy. So even though the colour is not changed, the momentum and energy are.

18.5.2 Annihilation and Creation

Quarks can be annihilated to produce photons, for example, but also gluons. When quarks annihilate into a photon, they must be each other’s anti-colours. Otherwise, colour is not conserved. But when they annihilate into a gluon, they can be of a different colour (see Fig. 18.19). For this diagram there are two colour stream diagrams on the right. They illustrate again the view from two different observers that move with different velocities. In one view the red quark is an anti-quark, while in the other it is not; and vice versa for the green quark. The gluon is $g\bar{r}$ or $r\bar{g}$ in both cases. So, we end up once again with a Lorentz invariant gluon $g\bar{r} + r\bar{g}$.

Within a hadron, there are many gluons between the quarks. These gluons can split into virtual quark–anti-quark pairs. Since gluons are massless, the production of such massive pairs is usually far from the resonance point. Hence, they will be quark disturbances and short-lived. These pairs recombine to produce gluons again. They

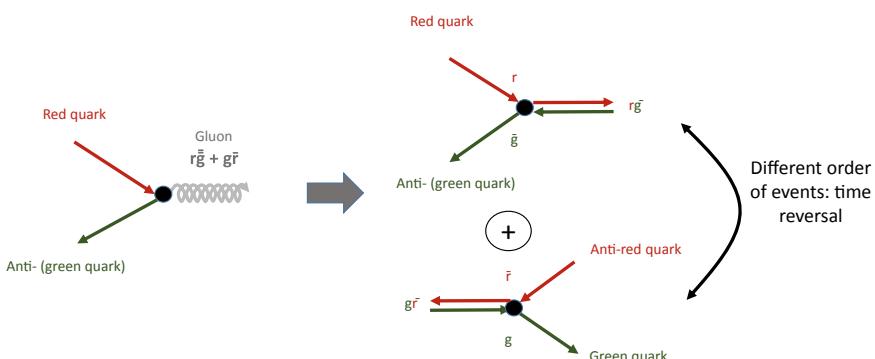


Fig. 18.19 Quark annihilation process

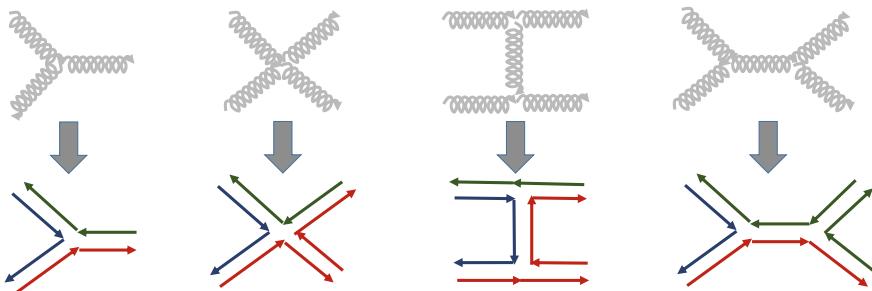


Fig. 18.20 Gluon–gluon interactions

must be gluons in order to preserve the colour in this process. So, between the quarks in a hadron, we find not only a sea of gluons, but also a sea of quark–anti-quark pairs. Just like the electron, the hadron consists of a whole bunch of disturbances: many waves in quark and gluon fields.

The annihilation of two quarks into a gluon is called gluon radiation, as a gluon gets radiated away. The opposite process, a gluon decaying into a quark–anti-quark pair is called gluon splitting.

18.5.3 Gluon–Gluon Interactions

Gluons carry colour and so they can interact with each other. This is an essential part of the strong force as it is responsible for asymptotic freedom and colour confinement. So, let's see what such interactions look like (see Fig. 18.20) [Ref. 23, handout 8; Ref. 30].

The examples show that gluons can interact with three gluons at the same time or with four at the same time. The colour streams show that colour is conserved in all cases. From these diagrams we see that the colours of a resulting gluon must be determined by those of the incoming gluons.

18.5.4 Proton–Anti-proton Collisions

There are several accelerators in the world that smash protons and anti-protons together in order to research elementary particles, e.g., LHC (CERN) or Tevatron (Fermi lab). In these collisions, the individual quarks of the (anti-) protons interact with each other, and even the gluons of the (anti-) protons can interact. Such interactions usually lead to jets of new particles coming out in two or three directions, as discussed before. So, let's see what types of interaction play a role in such processes. In such collisions, it is important to note that the momentum of the proton is carried

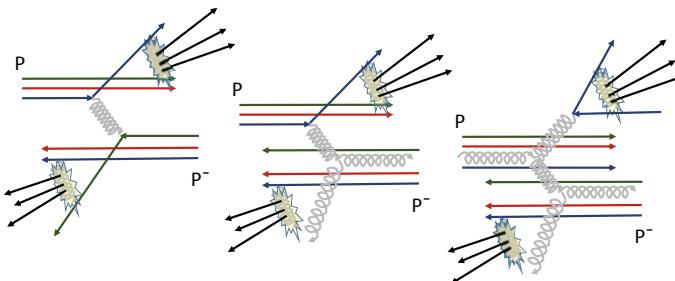


Fig. 18.21 Jet production as a consequence of quark-gluon interactions in a proton-anti-proton collision. (1) Momentum transfer ejecting two quarks out of the (anti)proton. (2) A quark being ejected and a gluon interacting with another gluon producing a third gluon that generates the jet. (3) Only gluons in the proton and antiproton interact. One gluon produces a jet. The other splits into a quark-anti-quark pair that produces the jet

not only by the quarks, but also by the gluons. All parts of the proton carry the momentum of the whole proton. When the momentum is distributed unevenly over the parts in the proton, it will make the proton stretch. When that happens, the quarks get further apart. The strong force will slow them down and pull them back. The result is an oscillation of quarks inside the proton. It is the average momentum that makes the total momentum of the proton. The oscillation itself makes up part of the mass energy of the proton. When the oscillation is too strong, a quark can free itself and produce a quark-anti-quark pair. This can happen regularly, as we will see when we discuss the force between protons and neutrons in a nucleus.

When a proton smashes into an anti-proton there are many things that can happen. There can be annihilation of quarks or the entire proton, production of electromagnetic radiation, production of Z and W bosons, etc. Let's focus on the quark-gluon interactions and the production of jets.

Although gluons usually create an attractive force, in high velocity collisions, momentum can be transferred between quarks that will make them scatter out of the original (anti)proton. When they do, they will each create a jet and there will be two jets coming from the collision. Alternatively, gluons can interact with each other and this can also lead to jets. Some examples are given in Fig. 18.21.

18.5.5 Residual Strong Force or “Nuclear Force”

We saw before, that white particles cannot interact strongly. Yet we see that the strong interaction binds protons and neutrons (together called nucleons) into one nucleus. Protons in particular would otherwise not get together since their electric charges repel each other. So how does this work?

We have seen that when a quark has enough velocity to get a little further away, the gluon band produces a new quark-anti-quark pair. This can happen spontaneously

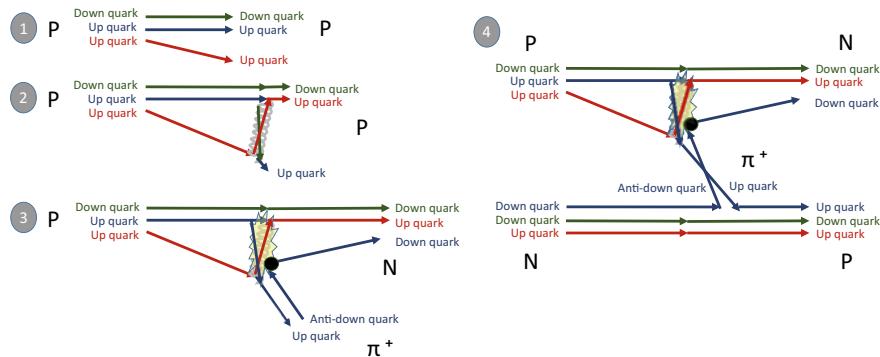


Fig. 18.22 Meson exchange in the nucleus. (Step 1) One of the quarks in the proton moves further out. (Step 2) Gluons appear between the quark and the other quarks of the proton, exchanging colours. (Step 3) The gluons break up to produce a down–anti-down pair. The anti-down pairs with the up to form a pion (π^+). The down pairs with the two other quarks of the proton and becomes a neutron. (Step 4) The pion is captured by a neutron that annihilates a down against the anti-down of the pion. It takes in the up of the pion. Hence, it changes into a proton

in a nucleus. For instance, an up quark in the proton (uud) moves out of the proton and produces a pair. Let's say that the resulting pair is a down quark + anti-down quark. The anti-down quark teams up with the up quark coming from the proton (see Fig. 18.22). Consequently, it becomes a π^+ meson (pion). The down quark that was produced becomes part of the original proton. That now has two down quarks and one up quark (udd). Consequently, it has become a neutron.

The pion and the neutron are both white and can move away from each other. A little later, the pion gets absorbed by a neutron (udd). The anti-down quark of the pion annihilates with a down quark from the neutron and the up quark in the pion becomes part of the original neutron that has now become a proton (uud).

In the process, neutrons and protons may or may not get changed into each other. It resembles a bit the process of colour exchange between quarks. So, we see that a neutron is never long a neutron and a proton is never long a proton within a nucleus. Outside a nucleus, the meson gets reabsorbed and a neutron or proton never changes into the other. Consequently, a neutron remains a neutron and has time to decay outside a nucleus. Inside a nucleus the neutron changes into a proton before it can decay. However, it has more chance of decaying when it remains a neutron for longer. This can happen in a nucleus when there are too many neutrons in the nucleus. This will reduce the probability of changing into a proton and will make neutrons remain as neutrons for longer. Hence, the more neutrons in an isotope of some type of atom, the less stable it tends to get and the higher the probability that it will decay radioactively.

The meson that is produced by a nucleon is usually a pion. A pion consists of up and down quarks, the lightest quarks, which are also the building blocks of nucleons. Figure 18.22 illustrates that in general the meson should contain quarks from the original nucleon. Hence, it must be an up or down quark. The pion can also be a $d\bar{d}$

or $u\bar{u}$ (π^0) that does not change the original nucleon. A third option is $d\bar{u}$ (π^-), which changes a neutron into a proton and gets absorbed by a proton that changes into a neutron. The $u\bar{u}$ and $d\bar{d}$ combinations (i.e., π^0) are less stable since quark–anti-quark pairs can easily annihilate each other.

The effect of meson exchange is that the mass potential of the nucleon becomes less strong (the nucleon mass is reduced). Hence, the mass-springs in the nucleon wave become less strong. Put differently, the intrinsic energy (or “potential energy”) of the nucleon has been reduced by creation of the pion.

The closer two nucleons get, the stronger this effect since more mesons can be exchanged. A virtual pion (a pion disturbance) does not have a long life so the strength of pion exchange falls quickly when the nucleons are further apart.

We can compare this situation with a field strength. The meson “field” is stronger when nucleons are closer together. The stronger the (attractive) field, the lower the mass of the nucleon as a whole (we are talking about springs that are stretched, so the closer we get, the less potential there is in the spring). The lower the mass of the nucleons, the lower their frequency. The pion was produced from the velocity of a quark and the band of gluons that resulted from that. Hence, the pion production must primarily reduce the average quark oscillation velocity. This means that the frequency has been reduced, but the average internal velocity has been reduced even more by the creation of the pion. Remember that the relation between the frequency, velocity, and wavelength of a wave is given by $f = v/\lambda$. Hence, the effect must be that the average wavelength of the nucleon gets shortened every time a pion is produced. Consequently, when two nucleons are closer together, the effect is stronger and their average wavelength gets shorter. We have seen that a shortening of the wavelength when two objects get closer together has the effect of an attractive force (see Sect. 9.4).

You may wonder what the wavelength of a nucleon entails. Basically, a nucleon is a wave packet made of a combination of quark waves and gluon waves. The average of these waves can be summarized as the nucleon wave.

The effective force has for protons a maximum around a distance of 1 fm (10^{-15} m). Below 0.7 fm, two protons most strongly repel each other as a consequence of the fact that they are fermions that cannot be in the same position. Above 1 fm, the force gets weaker, but it is stronger than the repulsive force between the two electric proton charges until about 1.7 fm, about twice the size of a proton. Beyond that distance the electromagnetic force wins since it has a much longer range than the meson driven nuclear force. Consequently, this gives a range of 0.7–1.7 fm over which protons can attract each other. Two protons bind better when their spins are opposite, since then they can get closer together without excluding each other.

For neutrons, the lower bound exists as well, as they too cannot be in the same position. But since they have no electric charge, they will also attract each other beyond 1.7 fm. This makes it possible for neutrons to be a kind of “glue” between protons. As an example, an atom that consists of two protons alone (no neutrons) is highly unstable. This is an isotope of helium, also called a diproton. The protons do attract each other, but have to remain at a certain distance as well as have opposite spins. This just about gives the diproton a negative binding energy. So, it can’t last. Consequently, an atom that has too many protons is not stable either. When more

nucleons get together in a nucleus, relatively more neutrons are needed to glue it together. At some point too many neutrons are needed to keep it together and neutrons have a high probability of decaying. This is why the heavier atoms get, the more unstable they are.

The nuclear force is also called the residual strong force since it derives from strong interactions between quarks and gluons by the creation of mesons that carry the nuclear force. The meson exchange is in principle not produced by a local symmetry. Hence, they are not a new field that is created to propagate away a rotation or phase shift. The mesons that are exchanged are themselves composite particles. However, the idea that protons and neutrons can change into each other can be considered as a symmetry, with mesons being produced when they turn into each other. The funny thing is that the effect is quite similar and can be treated mathematically in the same way. The exchange of mesons can be described by the same type of potential, known as the Yukawa potential for the strong force.

There are some other effects relating to pion production by protons and neutrons. For instance, a neutron has a magnetic moment. This is interesting, since it does not have a charge. However, it can produce a pion and then reabsorb it. When it produces a π^- , it temporarily becomes a proton until it reabsorbs the pion. This situation is similar in some ways to a photon producing an electron–positron pair for a short time. One consequence is that the photon can be polarized, and so can the neutron when it briefly dissociates into a proton + pion. The neutron can have a magnetic moment originating from the charge polarization between the proton and pion.

18.6 Masses of Quarks, Mesons, and Baryons

We saw before that quarks are free above the Hagedorn temperature (2×10^{12} K). Below this temperature, quarks cannot be free and clump together in baryons and mesons. This is similar to moisture in the atmosphere, which clumps together to form raindrops below a certain temperature. So, the formation of baryons and mesons is in fact “condensation” of quarks and gluons into such formations. This condensation process can be seen as a breaking of symmetry. At the same time, it requires energy to lift the condensation and recreate the symmetric unbound situation. Consequently, there is binding energy in baryons and mesons.

So, let's look at the bare masses of the up and down quarks and compare them with the proton mass. The bare mass of a quark is also referred to as the “current quark mass”:

- Up quark bare mass: 2.3 MeV
- Down quark bare mass: 4.8 MeV.

A proton consists of two up quarks (= 4.6 MeV) and one down quark (4.8 MeV). In total this gives 9.4 MeV. However, the proton has a mass of about 938 MeV. So, the mass of the bare quarks makes up only about 1% of the total mass of the proton.

The bare mass can be estimated since quark masses do not go to infinity when we come closer, as a consequence of asymptotic freedom [Ref. 28].

However, the bare mass does not explain the mass of the proton. In order to understand that, we first need to look at the dressed mass of a quark, often referred to as the “constituent quark mass”. So, what should be that dressed mass? This turns out to be a whole lot more difficult to measure or estimate. After all, quarks do not appear as free particles in nature, unlike protons. So, we can measure the proton mass, but not the dressed quark mass. Several attempts have been made to estimate it using complicated models that we will not elaborate on here. One of the problems is that the dressed quark mass seems to depend a lot on the quark’s momentum in those models. Some results, however, suggest dressed masses for the lightest quarks of about 400 MeV [Ref. 28]. That would make sense. Added up, three quarks would have a mass of about 1200 MeV, which is some 300 MeV heavier than the proton. Consequently, the binding energy of the quarks in the proton would be of that order.

The models that are used support the idea of a quark being surrounded by a cloud of gluons, much as an electron is surrounded by virtual photons. The electron mass is influenced by that. In this case, the dressed quark mass is a lot higher than its bare mass. All that mass is produced by the energy content of the gluon cloud. When three such quarks + gluon clouds bind together into a proton, the binding energy is very large too. Hence, the proton is a lot less massive than the sum of the constituent quarks including their gluon cloud. Remember that the binding energy is equal to the energy needed to separate the constituents. That is hard to define as well in this case, since separation of quarks is not possible without producing new quark–anti-quark pairs. A quark trying to escape will have no difficulty producing a (virtual) meson, since the mass of a pion, for example, is only 140 MeV, well within the range of the estimated binding energy within the proton.

Another interesting matter is the mass of a pion. It is very light, while two dressed quarks would sum up to 800 MeV, the pion is only 140 MeV [Ref. 28]. That would imply an extremely high binding energy of 660 MeV. It is not clear why that is.

Summarizing, we can say that the masses of baryons and mesons are determined by the gluon cloud around quarks rather than by the bare quark mass. This becomes apparent only when quarks get bound into baryons and mesons. After all, it is only then that quarks and gluons are no longer free and show behaviour in which we have to add up the quark mass and gluon cloud mass into “dressed mass”. That’s why it is called the *constituent* quark mass, since it only applies to quarks that are constituents of hadrons. Consequently, the symmetry breaking process of condensation into hadrons is mainly responsible for the hadron mass.

18.7 Fundamental Particle Overview 4

We can now complete our overview of fundamental particles (see Fig. 18.23). The only thing we need to add is the SU(3) symmetry, leading to three colours for particles (red, green, and blue) and three anti-colours for anti-particles (resp. turquoise, purple,

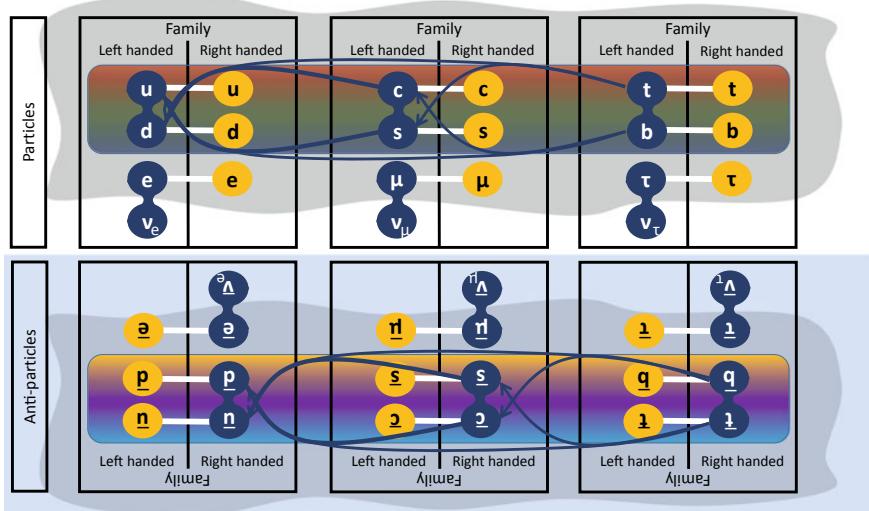


Fig. 18.23 Complete overview of fundamental particles and the gauge fields that cause them to interact

and yellow). The 8 gluon fields represent the rotations between those three colours and anti-colours.

Let's summarize what we have found. In total we have 6 quark flavours $\times 2$ chiralities $\times 3$ colours = 36 quarks. We have 6 left-handed leptons and 3 right-handed leptons. That makes a total of 45 fundamental particles. And of course, 45 anti-particles.

Each particle is an excitation of a type of field. Hence, 45 particles represent 45 different types of field. Many of these fields only differ by a rotation. The coloured fields turn into each other through an SU(3) symmetry operation. Left-handed fields can turn their upper face into their lower face through an SU(2) symmetry operation. Each field except the neutrinos can shift phase through a U(1) symmetry operation.

The symmetry operations are rotations that give rise to ripples in the vacuum, propagating the change away. The U(1) phase shift is propagated by the photon gauge wave. The SU(2) flavour rotation is propagated away by the W^+ , W^- , and Z^0 gauge waves. The SU(3) colour rotation is propagated away by 8 gluon gauge waves. Such gauge waves take a little potential out of the field waves that produced the rotation and propagate that away as well. The impact on other waves is that they get rotated by the gauge waves and absorb that potential. The effect of the potential on the wave can be that it either shortens the wavelength or extends it. When it shortens the wavelength, the paths of the waves get bent towards each other, thus producing an attractive force. When it extends the wavelength, the paths of the waves get bent away from each other, thus producing a repulsive force. The other effect of the potential is that it increases or decreases the “spring strength” or rather the elasticity in the field in an exactly opposite way to the change in momentum. The

result is a transfer of momentum and energy. The condition for this to happen is that the field strength is not homogeneous. The bending of a path is the result of the difference in shortening/extending the wavelength depending on the path the wave travels along.

So, there are $8 + 3 + 1 = 12$ gauge fields. The gauge bosons are the excitations of those fields. On top of these we have the Higgs field. The excitation of this field is the Higgs boson. This field interacts with all other fields, except gluons, photons, and possibly neutrinos. Since it got broken, it has pervaded the entire universe in a uniform way. Consequently, it acts as a constant spring towards the fields it interacts with. That means that excitations in those fields (the particles) must be given enough energy to overcome the Higgs springs. The strength of these springs depends on the strength of interaction with Higgs. This energy is what we call the bare rest mass of these fundamental field excitations (particles). The Higgs field also changes the wavelength, but since it is homogeneous (everywhere the same), this does not result in any bending of the paths of field excitations. Hence, no force.

The observed masses of the fundamental particles further depend on the cloud of disturbances (gauge waves) around them. Those gauge waves produce particle–anti-particle pairs that also impact the properties of the particles. Hence, the observed (dressed) mass of a particle is a sum of the bare mass and the cloud of disturbances around it.

All the interaction types produce their own springs, leading not only to a force, but also to a type of mass we call potential energy (when attracting particles are far apart) or binding energy (when they are bound). The latter represents the missing mass compared to when we add up the masses of the constituent particles.

In a nutshell this is how the standard model works. It explains the fundamental particles and forces we know about as well as most of their properties. Over the past decades it has been tested extensively in many particle accelerators, cosmic ray experiments, and laboratories that have done research in this field [e.g., Ref. 23, handout 14; Ref. 60, p.196; Ref. 30, p. 641]. The results confirm the picture we have sketched in this book.

Personally, I consider it a fantastic achievement of the literary tens of thousands of physicists that have been calculating and experimenting to arrive at this model, as well as the numerous technicians, mathematicians, programmers, and many others that have given them the tools to do this. Consequently, we can consider this model as one of the great achievements and insights of humankind. For the first time in history, we are able to understand at some level what the universe is made of and how forces come about. Until about 100 years ago, we could only understand that many phenomena are a consequence of particular types of force (gravity, electromagnetism), but we could not understand how such “forces acting at a distance” could come about. And now we can.

Chapter 19

Gravity as a Field



We have seen how all matter and forces are built up from waves. These waves use the vacuum as a medium to wave in. However, we have not said anything about gravity yet. Einstein's general theory of relativity explains gravity from a geometrical perspective. At first sight this may look completely different from what we have been discussing so far. The fact that physicists have tried and as yet failed to combine quantum field theory (QFT) with general relativity (GRT) is often perceived as a sign that they are completely on the other side of town. If that is your impression as well, you are in for a treat.

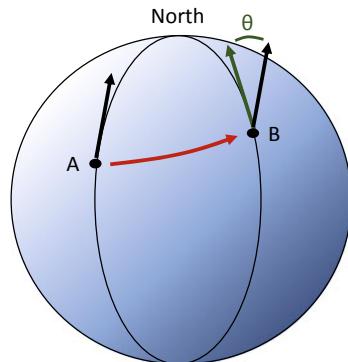
19.1 A Field Theory of Gravity

In fact, gravity has been described as a field theory. So, let's see what that looks like and how that compares to GRT [Ref. 30, pp. 403–407].

First of all, if we want to describe gravity as a gauge wave, we need to find a symmetry. That symmetry would need to apply to all energy since all energy interacts via gravity. There is one symmetry that applies to all energy: the Lorentz transformation. All laws, waves, and their energy must transform under Lorentz transformations. This is a global symmetry. What if we made it into a local symmetry just as we did with the other symmetries?

Let's start by looking at a coordinate translation in space and time. What is the effect of a coordinate translation? Take for example a sphere (see Fig. 19.1). Pick a random spot A on the sphere and create a vector that is pointing north. Now suppose we translate that vector to point B on the sphere, but make sure that the vector does not change direction. Now draw a new vector at point B that is pointing north. You will see that the original vector makes an angle with the new vector. So, the translation caused some kind of rotation to happen even though we did not rotate the vector! Such a rotation implies a change of physics. For example, when the arrow is the direction of a magnetic field, then the magnetic field will have changed direction as

Fig. 19.1 A coordinate transformation on a sphere. The vector at A points north. When a coordinate transformation moves it to B without changing its direction, it no longer points north. A vector pointing north at point B (green) makes an angle Θ with the original vector pointing north at A (black)



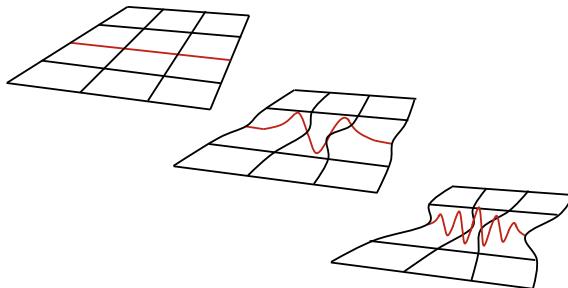
a consequence of a simple translation in space. Of course, this would not happen in flat space. But when space–time is curved, a translation can change the physics. So, in order to do a translation in curved space–time without changing the physics, we must rotate the vector as well. Consequently, either the vector does not change, but then the physics has changed, or the vector rotates so that it points north in its new position and the physics has not changed.

A Lorentz transformation is in fact a coordinate transformation, such as changing to a different frame of reference (e.g., another speed). That is not the same as a coordinate *translation*. However, it too has a geometric character. So, for argument's sake, we can use the example above as a comparison. When Lorentz symmetry is made local, this means that locally the physics does not change under a Lorentz transformation, and this in turn means rotating the vector. In order not to change the physics elsewhere, the transformation must be propagated through space–time. Consequently, the character of that propagated transformation is that it rotates the vector so that the physics does not change. It can only rotate the vector by curving space–time. This curving is similar to the curve of the sphere. Consequently, making Lorentz symmetry a local symmetry leads to gauge waves that curve space–time.

The Lorentz symmetry applies to all types of energy in the same manner. So, if it is made local, that will be the same for all energy as well. The consequence is that all energy “couples” to the gauge waves with the same strength per unit energy. So unlike U(1) symmetry, where some waves tend to shift phase and others do not (or with a different strength), there is no difference between types of waves with respect to the Lorentz symmetry. All transform coordinates the same way, and hence, all produce the same gauge wave per unit of energy. Consequently, gravity is universal. This also means that unlike the U(1) symmetry, gravity does not need a coupling strength. Or you might say that the coupling strength is determined directly by the mass-equivalent of the energy.

One way to view this is to take a wave and look at what it does to the medium in which it waves. Basically, it wrinkles that medium. That means that it stretches the medium but also contracts the “lines” of the medium into a small area (see Fig. 19.2). As the frequency goes up, more curves in the wave contract the medium

Fig. 19.2 Waves curving the vacuum. On the left the vacuum is flat as there is no wave. In the middle there is a wave with a low frequency. It contracts the vacuum, which becomes curved. On the right a high frequency wave contracts the vacuum more strongly, leading to more curvature



more and more. Suppose the medium is the vacuum and we view its contraction as the curvature of space-time. Could that be an accurate picture? Well, in GRT it is energy that determines the curvature of space-time. If we are a bit naughty and we take as energy the quantum mechanical definition $E = hF$, we find that it is the frequency F that determines the curvature of space-time. So, the idea is not that bad. Of course, in this picture we assume that the vacuum is such a stretchable medium. However, we have no idea what the vacuum really is. Still, it gives a vivid picture to help us understand how waves could curve space-time.

Mass is the official “charge” of gravity. But from this discussion we can conclude that the frequency of the wave can be seen as the “charge” of gravity. You may be wondering how this works. After all, mass was created by Higgs, wasn’t it? So how does that relate to frequency? Well, Higgs required us to put extra energy into the wave. The extra energy represents the mass. Higgs provides the springs to the wave to contain that extra energy. The extra energy does the same job as an increase in frequency (remember that we were adding mass frequency to kinetic frequency in the section on energy and mass).

So, we have been doing some speculation here and our conclusion is that space-time gets curved by a wave. In turn, a second wave (representing another particle) follows the path of minimal action according to the path integral. The path of that wave will be bent in line with the curvature. So, the curvature leads to the gravitational force. How strong is that force? Obviously, the higher the frequency, the greater the curvature, the stronger the force. So, the force must depend on the frequency of both waves (i.e., the mass of both waves). Moreover, it will decline with the square of the distance. That looks like Newton’s law of gravity.

However, there is another effect that depends on the elasticity of the vacuum. Suppose the vacuum has a very high elasticity, like a cloth lying loosely on the table. We can easily create a wave in it, but it will not curve much. If we had a tightly stretched rubber sheet, it would be much more difficult to create a wave, but we can imagine that it will stretch much further as well. In order to include this stretch, which depends on the elasticity of the vacuum, we need another factor. That factor is the gravitational constant. So, the gravitational constant depends on the elasticity of the vacuum. The less elastic the vacuum (think about the tightly stretched rubber sheet), the greater the gravitational constant and the stronger the force of gravity. With these speculations, we obtain the full formula of gravity according to Newton:

$$F = G m_1 m_2 / r^2$$

We saw before that the speed of light C depends on the elasticity of the vacuum: the less elastic it is, the higher the value of C . We can argue the same for another important constant: Planck's constant. This constant appears in the relation between energy and frequency: $E = h\nu$. When a medium is not elastic (very stiff, e.g., a tightly stretched rubber sheet), it would take a lot of energy to create a wave in it. When it is more elastic (e.g., a tablecloth lying loosely on the table), it hardly takes any energy to create a wave in it of the same frequency. Consequently, we find that h decreases with the elasticity.

So now we have three fundamental constants of nature that depend on the elasticity of the vacuum: c , h , and G . All three tend to increase with decreasing elasticity. However, their dependence on that elasticity is different. This is because they describe different consequences of the elasticity. The velocity of waves is a different thing than the energy you need to create a certain frequency. And these are different things compared to the stretch a wave creates in the vacuum. We can play around a little with these constants and combine them in such a way that we get a length out: this is the Planck length. It is defined as

$$\text{Planck length} = \sqrt{hG/c^3} = 10^{-35} \text{ m (approximately)}$$

What could this length mean? It could have something to do with the fundamental elasticity of the vacuum. Hence, it could say something about the length scale of the structure of the vacuum. There is an analogy with the distance between the atoms in a solid that carries phonon waves. That distance determines the elasticity in the solid as experienced by the phonons and it says something about the structure of the solid. In any case, it seems that, at the Planck length scale, gravity gets so strong that space-time would be all curled up. That is not a situation in which we could easily define fields, so it also suggests that we are entering the realm of the structure of the vacuum. All this is *speculative* though.

Let's get back to gravity as a gauge wave. When a particle transforms coordinates locally, the tension within space-time has already changed "here" and not yet "there", i.e., the rotation/translation has already taken place "here", but not yet "there". It is this difference in rotation/translation that creates the gauge potential. This potential is similar to the gauge field A in the electromagnetic field. The potential is propagated by a gauge wave. The gauge waves are called gravitons, just as electromagnetic gauge waves are called photons. Gravitons are probably massless. If we examine the way the interaction works, it turns out that it looks exactly like a coordinate transformation [Ref. 30, p. 137]. Hence, the interaction operates on another wave like a curvature of space-time. Summing up, mass tells space-time how to curve and (curved) space-time tells mass how to move [Ref. 76]. Consequently, it does look different from the potentials we are used to. The graviton has a more geometric character. However, one can recognize the gauge waves and the interaction with mass as a curvature of space-time.

Gravitons have not been detected. This is not so strange when we realise how weakly they couple to matter. We would need a detector the size of a planet in the neighbourhood of a black hole to detect one graviton per decade. So direct detection of gravitons is impossible. However, other experiments can help us get a picture of the characteristics of gravitons. Recently detected gravitational waves came from 130 million light years away and arrived at almost the same time as the radiation that was emitted by the same source [Refs. 53, 54]. A gravitational wave is basically a huge assembly of gravitons. So, the velocity of those waves indicates that gravitons are either massless, or have an extremely tiny mass.

The potential that is propagated contains energy, which leads to an interesting situation. When the gauge waves carry energy, they must curve space-time as well! So, gravitons are just like weak bosons and gluons: they produce gauge waves themselves. The effect of this was different for weak bosons and for gluons. The mass of the weak bosons limited the power of the force. Gluons, on the other hand, showed characteristics such as colour confinement and asymptotic freedom. So, what should we expect from the graviton?

On longer scales than the Planck scale (10^{-35} m) gravity is far too weak to produce asymptotic freedom. However, on smaller scales, it is stronger. The closer we get to a particle, the stronger the gravitation. Hence, there is more energy in the field. This means that it produces more gravitons. This effect will be enhanced more and more the closer we get to a particle. If we consider distances below the Planck scale, the gravitational force will go to infinity. Is this theory renormalizable? That is, can we find counter-terms that would overcome the infinity? Mathematically, below the Planck scale, we would need an infinite number of counter-terms to do that. Hence, the theory is not renormalizable. There is no screening effect that could do such a thing.

On the other hand, on longer scales the theory does produce the same results as general relativity. For instance, one can calculate the deviations in the orbit of Mercury using GRT and using QFT and end up with the same result [Ref. 30, p. 138]. On longer scales than the Planck length, the theory turns out to be in exact agreement with GRT. So, the biggest problem of QFT is the infinities below the Planck scale, a problem that GRT does not have. Otherwise, it is a perfectly viable theory.

When we look at the picture of waves contracting the vacuum, we can see that the curvature (stretch) is greater the closer we get to the wave. The greater the curvature, the more energy is in it and the more strongly it curves its environment. Going closer in, we can go to infinity. But this requires the wave to be point-like and clearly, it is not. So, this picture suggests something different. It suggests that there is contraction in the middle of the wave, but that it never goes to infinity. This picture is of course a speculative interpretation and not what the mathematics describes! But the drawback with the mathematics is that it too relies on assumptions. One of these assumptions is that the shape of the wave does not have to be included in describing the way it produces a coordinate transformation. The mathematics only describes an interaction potential. It does not take the structure of the wave/particle into account. At the Planck scale this assumption is no longer valid. So, we should start to take

a look at the structure of our waves. This would require a deeper theory that has validity on the Planck scale.

Another way of looking at it is that the structure of the *vacuum* (as opposed to the structure of the particles) probably starts to play a role at the Planck scale. As we saw before with waves in a crystal, we cannot define such waves on a scale smaller than the distance between the atoms in the crystal. So, the size of the structure of the medium determines the minimum size of a meaningful field theory. Hence, we could say that below the Planck scale field theories of waves in the vacuum become meaningless and we need another theory about the structure of the vacuum to understand what happens at that scale.

Concluding we can say that gravity can be defined as a field theory that gives the same results as GRT above the Planck scale. A *more speculative* discussion leads us to a picture of how space–time gets curved by waves. It is the frequency of waves that pulls the vacuum together and curves space–time. The higher the frequency, the stronger the curvature. Henceforth, we could connect the three fundamental constants of nature c , h , and G to the elasticity of the vacuum. Finally, we discussed the way the Planck length represents the scale of the universe at which space–time would get too strongly curved to define a meaningful field theory. Hence, we now need a theory of the structure of the vacuum to show us what to expect below the Planck length. This theory would have to determine the stage at which fields can be defined and would have to show how to calculate the elasticity of the vacuum that gives us the constants of nature.

19.2 Background Independence

Let's discuss the stage on which our theories play out. Take for instance QFT. In order to define waves properly, we need a flat time. For instance, we calculated the path integral using time to record the phase of the wave. Quantum states evolve against an independent background of time. Suppose that waves can curve space–time (as in the speculative picture we discussed in the previous section). We would get into trouble calculating all this in QFT. It is unclear how this should work.

Another problem is: what are gravitons made of, when they cannot change space–time in QFT? The results discussed before suggest that they must be made of space–time itself. The gravitational field is characterized by geometric attributes such as length, areas, and volumes. Therefore, quanta of the gravitational field should correspond to quantized lengths, areas, and volumes. Then there must be a relation between such geometric entities and energy in a quantized way, just as a quantum of the electromagnetic field should give rise to a frequency that equals an energy. So, we would expect a quantum of space–time to have a specific frequency and hence energy. However, if it has a frequency, it must be waving in something. But what would it be waving in if it is made of space–time itself? Gravitons seem to be waves *of* the background rather than *on* the background. However, gravitons do not say anything about the background itself. We can just say how they change the background as a

consequence of the presence of energy. Hence, gravitons would be waves of space and time in some way and other waves would not be. But then, what makes general waves connect to gravitons? After all, the basic idea was that all energy, hence all waves, create coordinate transformations that are propagated through space-time.

On the other hand, we have GRT [Refs. 75, 76]. GRT does not need a background. It describes space-time as dynamical variables that can connect to matter directly. The idea of relativity makes it essential that there should be no background against which we can measure position, time, or velocity. This role of time is very different from the one formulated in QM and QFT. It makes it impossible for time to be the independent background that QM and QFT need as a reference. GRT is a continuous theory, so it cannot handle quantization. The assumption that space-time is continuous implies also that it is infinitely divisible. The question is whether this can be true. It is more probable that space-time is also quantized and there is some smallest unit of space-time.

This is the core of the problem that QFT and GRT cannot be combined. What we would need is a theory about the structure of the vacuum that can accommodate both theories. It must provide a stage for waves to wave in, it must allow itself to be altered under a coordinate transformation, and it must provide the background-independent structure needed by GRT. Another reason not to be background dependent would be that we would need to know what that background *consists* of. For that we would need another background to place it in, and the next question would be: what is *that* background made of? If we need yet another background, we could go on and on like that and never get to the bottom of it. We would like to land upon a theory that does not need a background (a stage) to operate in. We need a theory that in fact *is* the stage. Hence, it should be derived from only a few, very basic “first principles”. So, here’s the puzzle that present-day theoretical physicists are facing.

19.3 Other Problems

There are also some other problems. First of all, QFT itself is not complete. There are at least 18 parameters we have to put in from measurement which cannot be derived from basic principles. These parameters fall into four categories:

1. Masses of particles: there are 6 quark masses, 3 lepton masses, and at least one Higgs mass whose values we cannot explain. Put differently, we cannot explain the size of the Higgs coupling to these particles.
2. Coupling constants: we cannot explain the strength of the 3 coupling constants of the electromagnetic field, the strong interaction, and the weak interaction.
3. Mixing angles: we cannot explain the 4 mixing angles of the quarks, e.g., why a top quark is sometimes a bit of an up quark or a charm quark.
4. And finally, we cannot derive the field strength of Higgs, also called the “vacuum expectation value” of the Higgs field.

Then there is the problem that there is more matter than anti-matter. Is the anti-matter to be found somewhere else? There is nothing that points that way. Suppose you have “anti-matter domains” and “matter domains”. Then, they would generate a lot of heavy radiation at the border between such domains. No such thing has been found so far. We have seen that breaking the symmetry could explain some imbalance between matter and anti-matter causing more matter to be present than anti-matter. But this effect is not nearly enough to explain the present-day situation in the universe.

Then there is dark matter and dark energy. This causes galaxies to rotate faster than would be expected on the basis of their mass content. Hence, there must be more mass, but what is that made of? In fact, only 4% of the energy in the universe appears to be made of the matter and energy we are familiar with. We are looking for unknown types of matter that do not (or hardly) interact with the matter we know. But other solutions are also being sought.

The hierarchy problem is a problem of parameters in physics that are related to each other, but differ *enormously* in magnitude. When parameters are related and influence each other, this can be well described if they are of similar orders of magnitude. But when these parameters differ by tens of orders, this becomes trickier. For instance, when you try to lift a washing machine using a lever, you can calculate how long the lever must be. This makes sense. But if the washing machine were 10^{24} times as heavy, the lever you calculate could easily be off by many orders of magnitude due to an extremely small mistake. So, what is the meaning of a theory that has to deal with such differences? The smallest change in one of the variables in such a theory would produce results that are way off. So, it becomes really hard to link parameters into one theory when these parameters differ by 10's of orders of magnitude.

For example, the mass of Higgs and the Planck mass differ by a factor 10^{17} . Another question is why is the weak force so much stronger than gravity, with a factor of the order of 10^{24} between the coupling constants?

The problem of the cosmological constant is in part also a hierarchy problem. The cosmological constant represents the average energy density in space-time. If we designed a theory at the Planck scale as a theory that could accommodate both GRT and QFT, we would expect the energy density in such a vacuum to be of the order of the Planck mass (or Planck energy). After all, any disturbance, wave, or fluctuation in such a structure at the Planck scale would have an energy of that order. However, the measured value of the cosmological constant is approximately a factor 10^{120} off. How could the average energy of any structure be 10^{120} times smaller than the energy a wave would carry in such a structure? This has been called “the worst theoretical prediction in the history of physics” [Ref. 75, p. 187]. Some process should exist that makes this average energy extremely small, but what could it be?

When we look at a theory of the structure of the vacuum, these complicated problems will have to be addressed. By now you will understand the extent of the puzzle that has to be solved.

There are several theories that try to address these problems by describing a structure of space-time or by avoiding the infinities in other ways. It would take us

too far of course to address these theories here, but there are excellent books giving an idea of how they work. Some of these theories are [Ref. 48] string theory [Refs. 36, 80], causal dynamical triangulations [Refs. 49–51], and loop quantum gravity [Ref. 47].

It turns out that it is hard to distinguish between such theories on the experimental level. Unfortunately, this is what is needed most, since only experiment can separate right from wrong.

QFT and GRT have both been checked to great detail in many experiments. Hence, these theories must represent some level of reality even though we cannot understand the structure of the vacuum that lies beneath. In this book I hope to have given a taste of what that reality might really look like.

Chapter 20

Further Reading



I will discuss two categories of material for further reading. The first category is popular science books. The second concerns what is available on the internet. I will not address the professional literature. So let's look at these two.

20.1 Pop Science

A lot of books have been written about particles and forces. In general, the attempt to keep things simple is admirable, but does come at a cost. Not all things are explained. In general, one may say that the more that gets explained, the more complicated the book. There is a kind of uncertainty principle for books about this subject: the better the explanation, the greater the complexity. Usually, books are either very complex (the professional literature) or leave out a lot (popular science). My book is subject to the same uncertainty relation. However, I chose a particular middle point in the sense that I describe more abstract concepts, but I still leave out the very complex math.

Let's look at some popular science books and what you may expect from them.

One of my favourites is QED by Richard Feynman [Ref. 32]. He focusses on explaining the path integral and interactions between charged particles. He does that very nicely and you will recognize the principle of adding phases, which I used in Sect. 9.3 in this book.

A good book that takes symmetry as a basis to explain the origin of forces is “The Force of Symmetry” by Icke [Ref. 35]. He uses the concept of the Chimera to explain how particles change face. He tries to get to the heart of the matter just as much as I do, but from a slightly different perspective. You will find that some things in this book are inspired by Icke, and some things are complementary.

A similar work that can be considered complementary is “Fearful Symmetry” by Zee [Ref. 33]. You will find similarities, e.g., in the way Zee explains the concept of action.

Then there are the immensely popular books by Hawking (e.g. [Ref. 36]). Some are well illustrated and cover a wide variety of subjects, including the history of the universe, gravity, and string theory. Some are explained to great depth while remaining on an understandable level. Some are only on the overview level. That sometimes makes the overall relation between things a bit abstract.

A different approach is taken by Brooks in “Fields of Colour” [Ref. 34]. He describes the world through the view of fields. In order to keep the different types of fields apart, he gives them a colour. These colours are not related to quark colours. Again, an inspirational book.

For the Dutch reader, there is “De bouwstenen van de schepping” by ‘t Hooft [Ref. 37]. An overview that gives yet other insights into the subject. It tells the story from a historical perspective using many anecdotes, explaining experiments and the problems theorists ran into.

Also from a historical perspective, “The Particle Odyssey” [Ref. 24] tells the story of the discovery of particles with fantastic illustrations of experimental results such as bubble chamber pictures including the explanation of the particle tracks.

You will find that these authors (and many others that I left out here) each have their own way to describe things and each adds different concepts and viewpoints that will help you to understand. Hence, reading more than one book will increase your understanding of the subject. Comparing the different ways of explaining stuff will do that too. I can only wish you a lot of fun in getting a grip on the way our universe works!

20.2 The Internet

The internet contains a lot of material, but not everything can be trusted to be correct. For example, there are a lot of fora where you find people who ask questions about scientific subjects. The answers are sometimes very good and sometimes they are not. The articles on Wikipedia also differ in quality. A good source are the universities. They often put lecture notes on the internet or whitepapers that can be very good. Scientific articles are often very detailed and hence not the easiest to follow. In this section I have made a selection of material I liked. Though you might find many rather technical and of course the material offered by universities belongs to the area of professional literature.

I found an interesting summary of group theory and the groups one finds in particle physics in “Applications of group theory to fundamental particle physics” by Bergan [Ref. 40]. Others are:

- “Group theory and physics” by Calvert [Ref. 43]. He has many more interesting subjects on his site.
- “Lie groups in physics” by Veltman, de Wit, and ‘t Hooft [Ref. 44]. A well readable and understandable introduction to group theory and its applications in particle physics, but mathematical.

- “An introduction to group theory” by Lin [Ref. 45] gives a vivid example calculation, but again it contains math.

For the understanding of waves, I found the lecture notes of Prof. Morin [Refs. 10, 11] very insightful. He explains well how to understand, e.g., group velocity and especially dispersion relations. He also wrote a clear explanation about the exchange between potential energy and kinetic energy (“The Lagrangian Method”) [Ref. 19]. Both are published as books as well. However, you need mathematics to follow these.

About the same subject, a very nice mathematical text is “A very short introduction to quantum field theory” by Prof. Stetz [Ref. 38]. You should read the part about the dispersion relation and the difference between the dispersion of massless and massive waves.

If you would like some additional material for understanding the path integral formalism, I found “Path integrals in quantum field theory” by Seahra a good read [Ref. 41].

I enjoyed some particularly clear handouts about particle physics from Prof. Thomson [Ref. 23]. He manages to put all the essentials, graphs, experiments, and explanations in a number of concise powerpoints. He covers the whole standard model and uses lucid examples. Although he covers quite a lot of math, there is also something to be found for non-mathematical readers.

It’s not only universities that have great material. Laboratories and facilities such as CERN also have a lot of good material. I especially enjoyed the layperson articles by Prof. Strassler [Refs. 12, 21, 22]. They provide a vivid story of virtual particles, the Higgs mechanism and quark jets.

A great site to visit is that of the Contemporary Physics Education Project (CPEP) [Ref. 39]:

The Contemporary Physics Education Project is a non-profit organization of teachers, educators, and physicists located around the world. CPEP materials present the current understanding of the fundamental nature of matter and energy, incorporating the major research findings of recent years.

There are many interesting chapters in the teacher’s guide, such as the one on symmetries and antimatter.

The video lectures of Richard Feynman can still be found on the internet [Ref. 13]. Although they are old, they are very worthwhile since Feynman had a very vivid way of explaining things. However, get ready for some math as well.

From a historic viewpoint, an interesting read is the original article by Einstein on the special theory of relativity. It is called “On the electrodynamics of moving bodies” [Ref. 42].

Finally, Wikipedia always offers an additional article you can read to get another viewpoint. Many articles go deep and quickly get very technical. It is especially amusing to look at the article “Spinors”, since it contains a nice computer animated film of a 720° turn before you get back to your starting point. For a more textual and lower entry level to some subjects, a good idea is to look at schools-wikipedia.org.

References and Sources

Quantum field theory and the standard model are a substantial subject that cannot be taken lightly. Very many theoretical and experimental physicists have been contributing to this. Altogether, these physicists have written a whole library full of books and articles about the subject. Writing a conceptual book about it would require proper reference to their work. However, over the past century (and more) so many ideas have been developed and built upon again and again that it seems impossible to do all of them justice.

Before writing this book, I studied the subject for many years, using many of those books and articles. During this time, I developed my own insights about how to explain certain aspects of the theory. Consequently, much of what I have written is original. On the other hand, many of the greatest ideas found an origin in the books I read.

First and foremost, there are five books that built my mathematical understanding of the subject. These are the books on QFT and the standard model by Lancaster and Blundell [Ref. 8], Zee [Ref. 9], Schwartz [Ref. 30], Stetz [Ref. 38], and Klauber [Ref. 29]. Lancaster and Blundell and Zee gave me the idea of springs attached to a rope, which I thankfully worked into a useful metaphor. These books also described how to calculate waves, second quantization, energy densities in a field, the energy-momentum tensor, symmetries, Noether principle and current, spinors, propagators, path integrals, QED, and renormalisation, to name just a few of the concepts discussed in this book. Most of what I wrote in Chaps. 5 till 11 I learned from those books.

For ideas about the path integral, I am especially indebted to the great Richard Feynman [Refs. 2, 13 and especially 32]. For an understanding of waves and coupled oscillators I leaned on the work of Prof. Morin [Refs. 10, 11] and Taylor [Ref. 5]. I used these ideas mostly in Chaps. 4, 5, 6 and 8. For understanding how potential energy and kinetic energy play out together, one needs to study the Lagrangian, which I learned from many books and articles [Refs. 5, 8, 9, 29, 30, 19, 61].

To understand QFT, one needs to have a handle on special relativity. For Chap. 6 and parts of Chap. 10 I based my discussion on the work by Woodhouse [Ref.

6], French [Ref. 31], and of course Einstein [Ref. 42]. Before that I learned about electrodynamics from Jackson [Ref. 3] and Griffiths [Ref. 4], which are used whenever I speak about the electromagnetic field throughout the book. Another basis in quantum mechanics I found in Shankar [Ref. 1] and Feynman [Ref. 2] which I used throughout the book as well, e.g., in describing the uncertainty relations, first quantization, (orbital) spin, the behaviour of bosons and fermions, and the quantum mechanical path integral. In this context, chirality and helicity play an important role. For that I would like to mention Lancaster and Blundell [Ref. 8] since this book seems to be most clear about the difference between the two concepts.

For the subject of symmetries and group theory I based my discussion on Jeevanjee [Ref. 7], Calvert [Ref. 43], Veltman, 't Hooft and de Wit [Ref. 44], van Suijlekom [Ref. 18], Bergan [Ref. 40], and Lin [Ref. 45]. This is a tough subject and I am greatly indebted to Icke [Ref. 35], Feynman [Ref. 32], and Zee [Ref. 33] for the conceptual ideas about symmetries and how they can produce a force. All this I applied and developed conceptually throughout the book and especially in Sects. 9.1, 9.2, 16.2, and 18.2. In describing how a gauge wave is produced I used the way they are calculated in the various books about QFT [Refs. 8, 9, 29, 30]. I used the same books to understand Feynman diagrams in Chap. 10 and Sects. 17.3 and 18.5. Virtual particles required the same sources, but there I would also like to mention the excellent site of Prof. Strassler [Ref. 12] which greatly helped me to develop my understanding of virtual particles in layman's terms. These subjects come together in the theory of QED for which one more mention of Feynman [Ref. 32] is in order.

To understand quantum decoherence I leaned a lot on Schlosshauer [Ref. 80] who wrote a comprehensive, not too mathematical, study on the subject.

With respect to the particle zoo, I am indebted to Close, Martin and Sutton [Ref. 24] for their great insight into the history of particle discoveries and their collection of pictures and tables.

For understanding the weak force, the Higgs mechanism, symmetry breaking, and CPT symmetry, I leaned on the more mathematical descriptions of Lancaster and Blundell [Ref. 8], Zee [Ref. 9], Schwartz [Ref. 30], van Suijlekom [Ref. 18], Jeevanjee [Ref. 7], and Thomson [Ref. 23]. For the practical and conceptual view, I looked at Icke [Ref. 35], Strassler [Ref. 21], and the CERN Higgs site [Ref. 15].

For quantum chromodynamics (QCD) I based myself on the more technical work by Schwartz [Ref. 30], van Suijlekom [Ref. 18], Jeevanjee [Ref. 7], and Thomson [Ref. 23, handout 8]. To complete that with conceptual insights I used Strassler [Ref. 22] and Icke [Ref. 35].

Gravity as a field is a short chapter that I based for the field part on the work by Schwartz [Ref. 30] and for the Lorentz transformation on Jeevanjee [Ref. 7]. For that I also used Hobson [Ref. 75], which I used in combination with Wheeler [Ref. 76] for the GRT part. The questions about background independence as well as other theories than string theory I took from various articles [Refs. 46–51].

In the second edition, I added a chapter about quantum decoherence. Although I used a variety of sources, I like to mention in particular the book by Schlosshauer [Ref. 80]. This book offers a great in-depth treatment of the subject, both qualitatively and quantitatively.

The second edition has been edited by Stephen Lyle. Since I am not a native English speaker, I particularly like to thank him for a great job improving the text in this book. His experience in physics proved extremely valuable as he understands the terminology very well.

1. Shankar, R. (1994). *Principles of quantum mechanics* (2nd ed.). Springer.
2. Feynman, R. (1977). *The Feynman lectures on physics* (Vols. I, II, III). Addison Wesley (or reprints of later date).
3. Jackson, J. D. (2009). *Classical electrodynamics* (3rd ed., C101).
4. Griffiths, D. J. (2015). *Introduction to electrodynamics* (4th ed.). Pearson Education.
5. Taylor, J. R. (2015). *Classical mechanics*. Univ Science Books.
6. Woodhouse, N. M. J. (2003). *Special relativity*. Springer.
7. Jeevanjee, N. (2015). *An introduction to tensors and group theory for physicists*. Springer.
8. Lancaster, T., & Blundell, S. J. (2014). *Quantum field theory for the gifted amateur*. Oxford University Press.
9. Zee, A. (2010). *Quantum field theory in a nutshell*. Princeton University Press.
10. Morin, D. (2010). Waves (Chap. 6). In *Dispersion*. Harvard University.
11. Morin, D. (2010). Waves (Chap. 10). In *Introduction to quantum mechanics*. Harvard University.
12. Strassler, M. (2011). *Virtual particles: What are they?* CERN. <https://profmattstrassler.com/articles-and-posts/particle-physics-basics/virtual-particles-what-are-they/>
13. Feynman, R. *Lectures by Richard Feynman*. <https://youtu.be/FF25Lwt73fg>
14. Costella, J. P., McKellar, B. H. J., & Rawlinson, A. A. (2001). *The Thomas rotation*.
15. CERN Higgs page. <https://home.cern/topics/higgs-boson>
16. Beringer, J. (2012). Particle data group. *Physical Review D*, 86, 010001.
17. ALEPH Collaboration, Decamp, D., Deschizeaux, B., Lees, J. P., Minard, M. N., Crespo, J. M., Delfino, M., Fernandez, E., Martinez, M., Miquel, R., Mir, M. L., Orteu, S., Pacheco, A., Perlas, J. A., Tubau, E., Catanesi, M. G., de Palma, M., Farilla, A., Iaselli, G., ... Zobernig, G. (1989). Determination of the number of light neutrino species. *Physics Letters B*, 231(4), 519.
18. Van Suijlekom, W. D. (2016). *Symmetrie in de deeltjesfysica*. Epsilon.
19. Morin, D. (2007). Classical mechanics (Chap. 6). In *The Lagrangian method*. Harvard University.
20. Beringer, J., Arguin, J.-F., Barnett, R. M., Copic, K., Dahl, O., Groom, D. E., Lin, C.-J., Lys, J., Murayama, H., Wohl, C. G., Yao, W.-M., Zyla, P. A., Amsler, C., Antonelli, M., Asner, D. M., Baer, H., Band, H. R., Basaglia, T., Bauer, C. W., et al. (2012). Review of particle physics: The CKM quark-mixing matrix (PDF). *Physical Review D*, 80(1), 1–1526 [162].

21. Strassler, M. (2011). *The known particles—If the Higgs field were zero.* CERN. <https://profmattstrassler.com/articles-and-posts/particle-physics-basics/the-known-apparently-elementary-particles/the-known-particles-if-the-higgs-field-were-zero/>
22. Strassler, M. (2011). *Jets: The manifestation of quarks and gluons.* CERN. <https://profmattstrassler.com/articles-and-posts/particle-physics-basics/the-known-apparently-elementary-particles/jets-the-manifestation-of-quarks-and-gluons/>
23. Thomson, M. A. (2009–2011). *Particle physics, handouts.* Department of Physics, High Energy Physics, Cambridge University. <https://www.hep.phy.cam.ac.uk/~thomson/partIIIparticles/welcome.html>
 - a. Handout 8: Quantum chromodynamics.
 - b. Handout 12: The CKM matrix and CP violation.
 - c. Handout 13: Electroweak unification and the W and Z bosons.
 - d. Handout 14: Precision tests of the standard model.
24. Close, F., Martin, M., & Sutton, C. (2002). *The particle odyssey.* Oxford University Press.
25. COBE data. Nasa. <https://lambda.gsfc.nasa.gov/product/cobe/>
26. Penta quark discovery, press release CERN. <http://press.cern/press-releases/2015/07/cerns-lhcb-experiment-reports-observation-exotic-pentaquark-particles>
27. Gell-Mann, M. (1995). *The quark and the jaguar: Adventures in the simple and the complex* (p. 180). Henry Holt and Co.
28. Roberts, C. D. (2010). *Exposing the dressed quark's mass.* Argonne National Laboratory.
29. Klauber, R. D. (2015). *Student-friendly quantum field theory.* Sandtrove Press.
30. Schwartz, M. D. (2014). *Quantum field theory and the standard model.* Cambridge University Press.
31. French, A. P. (1968). *Special relativity* (1st ed.). W. W. Norton & Company.
32. Feynman, R. P. (1985). *QED, the strange theory of light and matter.* Princeton University Press.
33. Zee, A. (2007). *Fearful symmetry.* Princeton University Press.
34. Brooks, R. A. (2016). *Fields of color: The theory that escaped Einstein.* Rodney A. Brooks.
35. Icke, V. (1997). *The force of symmetry.* Cambridge University Press.
36. Hawking, S. (2002). *The universe in a nutshell.* Stephen Hawking.
37. 't Hooft, G. (2013). *De bouwstenen van de schepping.* Bert Bakker.
38. Stetz, A. W. (2007). *A very short introduction to quantum field theory.* Oregon State University. Also published at freebookcentre.net
39. Contemporary Physics Education Project (CPEP). *Nuclear science, teachers guide.* cpepphysics.org
40. Bergan, W. (2015). *Applications of group theory to fundamental particle physics.* College of William and Mary.

41. Seahra, S. S. (2000). *Path integrals in quantum field theory*. Department of Physics, University of Waterloo.
42. Einstein, A. (1905). *On the electrodynamics of moving bodies*.
43. Calvert, J. B. *Group theory and physics*. University of Denver.
44. Veltman, M. J. G., 't Hooft, G., & de Wit, B. Q. P. J. (2007). *Lie groups in physics*. Institute for Theoretical Physics, Utrecht University.
45. Lin, H. H. (2010). *An introduction to group theory*. National Tsing Hua University.
46. Rovelli, C. (2003). A dialog on quantum gravity. *International Journal of Modern Physics D* 12(9), 1509.
47. Rovelli, C. (2003, November). Loop quantum gravity. *Physics World*.
48. Smolin, L. (2000). *Three roads to quantum gravity*. Oxford University Press.
49. Ambjorn, J., Jurkiewicz, J., & Loll, R. (2008, July). The self organizing quantum. *Scientific American*.
50. Loll, R. (2008). *The emergence of spacetime or quantum gravity on your desktop*. Institute for Theoretical Physics, Utrecht University.
51. Ambjorn, J., Jurkiewicz, J., & Loll, R. (2006). *The universe from scratch*. Niels Bohr Institute, Marc Kac Complex Systems Research Centre, Institute for Theoretical Physics.
52. The role of decoherence in quantum mechanics. *Stanford Encyclopedia of Philosophy*, April 2016.
53. Abbott, B. P., et al. (2017). GW170817: Observation of gravitational waves from a binary neutron star inspiral (PDF). *Physical Review Letters*, 119(16). LIGO Scientific Collaboration & Virgo Collaboration.
54. Wikipedia—GW170817.
55. Haroche, S., et al. (1996). Observing the progressive decoherence of the “meter” in a quantum measurement. *Physical Review Letters*, 77(24), 4887–4890.
56. Kaiser, D. (2006). Physics and Feynman diagrams. *American Scientist*, 93.
57. Bergan, W. (2015). *Applications of group theory to fundamental particle physics*. College of William and Mary.
58. Cottingham, W. N., & Greenwood, D. A. (2007). *An introduction to the standard model of particle physics* (2nd ed.). Cambridge University Press.
59. Weinberg, S. (1995). *The quantum theory of fields* (Vols. 1, 2 and 3). Cambridge University Press.
60. Peskin M. E., & Schroder D. V. (1995). *An introduction to quantum field theory*. Avalon Publishing.
61. Shifflett, J. A. (2015). *Standard model Lagrangian*.
62. Wu, C. S., Ambler, E., Hayward, R. W., Hoppes, D. D., & Hudson, R. P. (1957). Experimental test of parity conservation in beta decay. *Physical Review*, 105(4), 1413–1415.
63. Christenson, J. H., Cronin, J. W., Fitch, V. L., & Turlay, R. (1964). Evidence for the 2π decay of the K_2^0 meson. *Physical Review Letters*, 13, 138.

64. Couder, Y., Boudaoud, A., Protière, S., Moukhtar, J., & Fort, E. (2010). Walking droplets: A form of wave–particle duality at macroscopic level? *Europhysics News*, 41(1), 14–18.
65. Noether E. (1918). Invariante Variationsprobleme. *Nachrichten der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 1918, 235–257.
66. Thompson, W. J. (1994). *Angular momentum: An illustrated guide to rotational symmetries for physical systems* (Vol. 1, p. 5). Wiley.
67. Jaffe, R. L. (2005). The Casimir effect and the quantum vacuum. *Physical Review D*, 72.
68. National Institute of Standards and Technology. <https://physics.nist.gov/cuu/index.html>
69. Lenz, W. (1924). Über den Bewegungsverlauf und Quantenzustände der gestörten Keplerbewegung. *Zeitschrift für Physik*, 24.
70. <https://home.cern/science/accelerators/large-hadron-collider>
71. Cabibbo, N. (1963). Unitary symmetry and leptonic decays. *Physical Review Letters*, 10(12), 531–533.
72. Schwinger, J. (1951). The theory of quantized fields I. *Physical Review*, 82(6), 914–927.
73. Sauvage G. (1991). LEP results: Measurement of the Z^0 line shape in hadrons and leptons and of the lepton forward–backward asymmetries. In M. Lévy, J. L. Basdevant, M. Jacob, D. Speiser, J. Weyers, & R. Gastmans (Eds.), Z^0 Physics. *NATO ASI Series (Series B: Physics)* (Vol. 261). Springer.
74. Griffiths, D. (1987). *Introduction to elementary particles*. Wiley.
75. Hobson, M. P., Efstathiou, G. P., & Lasenby A. N. (2006). *General relativity: An introduction for physicists* (Reprint ed.). Cambridge University Press.
76. Wheeler, J. A. (1990). *A journey into gravity and spacetime*. Scientific American Library.
77. Brandelik, R., et al. (TASSO Collaboration). (1979). Evidence for planar events in $e^+ e^-$ annihilation at high energies. *Physics Letters B*, 86, 243–249.
78. Alexander, G., et al. (OPAL Collaboration). (1991). Measurement of three-jet distributions sensitive to the gluon spin in $e^+ e^-$ annihilations at $\sqrt{s} = 91$ GeV. *Zeitschrift für Physik C*, 52(4), 543.
79. Cartlidge, E. (2011, June 23). Quarks break free at two trillion degrees. *Physics World*.
80. Schlosshauer, M. (2008). *Decoherence and the quantum-to-classical transition*. Springer. Corrected 2nd printing.
81. Ramos, R., Spierings, D., Racicot, I., et al. (2020). Measurement of the time spent by a tunnelling atom within the barrier region. *Nature*, 583, 529–532.
82. Cramer, J. G. (2016). *The quantum handshake, entanglement, nonlocality and transactions*. Springer.
83. Easttom, C. (2021). *Quantum computing fundamentals* (1st ed.). Addison-Wesley Professional.
84. Tabb, M., Gawrylewski, A., & DelViscio, J. (2021, July 7). Quantum computers decoded (YouTube). *Scientific American*. <https://youtu.be/uLnGp1WTNFQ>

85. Bilenky, S. (2016). Neutrino oscillations: From a historical perspective to the present status. *Nuclear Physics B*, 908, 2–13.
86. Bilenky, S. (2010). Introduction to the physics of massive and mixed neutrinos. In *Lecture Notes in Physics* (Vol. 817). Springer.
87. Tetra quark discovery, press release CERN. <https://cerncourier.com/a/new-tetraquark-a-whisker-away-from-stability/>
88. Penta quark discovery, press release CERN. <https://home.cern/news/news/physics/lhcb-experiment-discovers-new-pentaquark>
89. Klimenko, A. Y. (2021). On the effect of decoherence on quantum tunnelling. *SN Applied Sciences*.
90. Pompili, M., Hermans, S. L. N., Baier, S., Beukers, H. K. C., Humphreys, P. C., Schouten, R. N., Vermeulen, R. F. L., Tiggelman, M. J., dos Santos Martins, L., Dirkse, B., Wehner, S., & Hanson, R. (2021). Realization of a multinode quantum network of remote solid-state qubits. *Science*, 372(6539).
91. Hensen, B., Bernien, H., Dréau, A. E., Reiserer, A., Kalb, N., Blok, M. S., Ruitenberg, J., Vermeulen, R. F. L., Schouten, R. N., Abellán, C., Amaya, W., Pruneri, V., Mitchell, M. W., Markham, M., Twitchen, D. J., Elkouss, D., Wehner, S., Taminiau, T. H., & Hanson, R. (2015). Loophole-free Bell inequality violation using electron spins separated by 1.3 kms. *Nature*, 526(7575), 682–686.
92. Aspect, A., Dalibard, J., & Roger, G. (1982). Experimental test of Bell's inequalities using time-varying analyzers, *Physical Review Letters*.

Index

A

Abelian, 102, 237
Absorption, 89
Acceleration, 56
Accelerator, 219
Action, 144
Alvaraz, 219
Anderson, 218
Annihilation, 125, 259, 261, 264, 312
Anti-colour, 290
Anti-matter, 149, 279, 328
Anti-particle, 125, 126
Anti-screening, 305
Anti-symmetric, 193, 289
Anti-symmetric wave function, 211
Armenteros, 219
Arrow of time, 148
Asymmetric mix, 253
Asymptotic freedom, 305
Attractive, 300
Attractive force, 115, 143

B

B, 220
Background, 326
Background independence, 326
Background radiation, 230
Baldo-Ceolin, 219
Balinese belly dancer, 200
Bankcard, 133
Bare charge, 157
Bare mass, 155, 317
Barnes, 219
Baryons, 221, 308
B-bosons, 232

Beta-decay, 276

Big bang, 230
Bjorkland, 218
Black body radiation, 249
Bloch sphere, 173
Boost, 185
Born rule, 30, 168, 192
Bosons, 179, 195, 221, 258, 288
Bottom, 220
Breaking of symmetry, 317
Bruno Pontecorvo, 271
Buckminster fullerene, 165
Butler, 218

C

Cabibbo angle, 271
Cabibbo rotation, 269, 273
Canonical quantization, 89
Carbon, 267
Carbon-date, 268
Casimir effect, 146
Cathode ray tube, 218
Causal dynamical triangulations, 16, 159, 329
Causality, 16, 148
CERN, 221, 225, 255, 284, 305, 306, 313
Chadwick, 218
Charge, 104, 254, 258
Charge conjugation, 278
Charge conservation, 214
Charged current, 263
Charge distribution, 157
Charm, 220
Chien-Shiung Wu, 276
Chirality, 179, 188, 202, 207, 274, 282

Christiaan Huygens, 3
 Classical mechanics, 9
 Classical observables, 170
 Cobe satellite, 230
 Coherent, 165
 Coherent field, 196
 Collapse of the wave function, 90, 123, 162, 167, 172
 Collisions, 221, 313
 Colour, 98, 289, 309
 Colour confinement, 300
 Colour hypercharge, 294
 Colour isospin, 294
 Colourless, 294
 Colour streams, 310
 Composite particle, 211
 Compton, 218
 Compton length, 37
 Condensate, 247
 Condensation, 317
 Conservation laws, 80, 215, 281
 Constant field, 116
 Constituent quark mass, 318
 Coordinate transformation, 322
 Coordinate translation, 321
 Cork, 219
 Cosmic radiation, 218
 Cosmological constant, 328
 Counter-terms, 154, 325
 Coupled oscillators, 85
 Coupling constant, 100, 159, 232, 235, 253, 327
 Coupling potential, 100
 Coupling strength, 159, 276, 322
 Cowan, 219
 CP symmetry, 278
 CPT symmetry, 128, 276
 CPT theorem, 281
 Creation, 126, 259, 261, 264, 312
 Crystal, 86, 87, 155, 159
 C-symmetry, 278
 Current quark mass, 317
 Curving space-time, 322
 Cyclotron, 218

D

D, 220
 Dark energy, 328
 Dark matter, 328
 de Broglie, 21
 Decay, 262, 283, 315
 Decay rate, 130

Decuplet, 308
 Degenerate, 181
 Degree of freedom, 189, 237
 De-localized, 211
 Delta, 287
 DESY, 220, 305
 Diproton, 316
 Dirac's belt trick, 200
 Dispersion, 39
 Dispersion relation, 39, 55, 57, 122
 Distance, 63
 Disturbance, 119, 129, 134, 146, 310
 Double slit experiment, 4, 34, 161, 166
 Down, 220
 Down quark, 231, 261, 266, 278, 288
 Dressed charge, 158
 Dressed mass, 155, 318

E

Early universe, 229
 Eigenfunctions, 86
 Eigenstate, 168
 Einstein, 3, 24, 49, 164, 321
 Elasticity, 42, 50, 324
 Elastic scattering, 258
 Electric charge, 101
 Electric field, 101, 143
 Electromagnetic field, 103, 240
 Electromagnetic force, 277
 Electromagnetic interaction, 143, 259
 Electromagnetic potential, 101
 Electromagnetism, 101
 Electron, 119, 125, 135, 138, 153, 218, 231, 261, 275
 Electron field, 99
 Electroweak force, 229
 Electroweak symmetry, 231
 Elliptical orbits, 181
 Emmy Noether, 84
 Energy, 9
 Energy density, 79
 Energy dissipation, 167, 171
 Energy flux, 81
 Energy peak, 284
 Entanglement, 163, 171, 173
 Environment, 165, 168
 Environment-induced superselection, 170
 Ether, 24
 Excitation, 77, 129, 251
 Expanding universe, 230
 External lines, 124

F

Face, 96, 235, 259, 261, 263, 292
Face changing, 96
Families of particles, 223
Fermat's principle, 45
Fermi lab, 221, 313
Fermions, 179, 194, 258, 274, 288
Feynman–Stueckelberg interpretation, 125
Feynman diagrams, 124, 144, 259
Feynman propagator, 133
Field, 19
Field amplitude, 74, 88
Field quantum, 76, 120
Field strength, 30, 94, 96, 100, 107, 139, 143, 184, 192, 195, 316, 327
Fine structure constant, 159
First principles, 327
First quantization, 71
Flatworld, 96
Flip chirality, 205
Fluctuating fields, 146
Flux, 81
Flux tube, 301
Force, 106, 110, 138, 196, 259, 263, 300, 316
Force working over a distance, 118
Fowler, 219
Free propagator, 129
Frequency, 9
Full propagator, 129
Fundamental particle overview, 224, 244, 284, 318
Fusion, 269

G

Gauge bosons, 100, 221
Gauge field, 100, 232
Gauge invariance, 104
Gauge theory, 101
Gauge wave, 122, 138, 232, 235, 259, 292, 322
Gell-Mann, 220
General theory of relativity, 321
Global phase change, 98
Global symmetry, 235, 292, 321
Glueballs, 301
Gluon, 220, 292, 296, 309, 318
Gluon cloud, 318
Gluon radiation, 313
Gluon splitting, 313
Goldhaber, 220
Goldstone boson, 251

Goldstone mode, 251

Gradient, 116
Gravitational constant, 323
Gravitons, 324
Gravity, 159, 321
Ground state, 75
Group, 93, 187, 235, 295
Group theory, 93
Group velocity, 38

H

Hadrons, 222, 288
Hagedorn temperature, 306, 317
Half-life, 267
Harmonic oscillator, 41, 73, 86, 251
Hawking, 16
 He^4 , 210
Heat capacity, 87
Helicity, 179, 201, 207
Helix, 187
Hendrik Casimir, 147
Hierarchy problem, 328
Higgs, 69, 116, 221, 231, 240, 249, 252, 274
Higgs boson, 253
Higgs field, 51, 73
Hooke's law, 73
Hypercharge, 254
Hypercharge current, 240

I

Igor Tamm, 77
Imaginary mass, 123
Inelastic scattering, 258
Inertia, 54, 56, 114
Infinity, 157, 306
Infinity problem, 153
Influx/outflux, 83
Inhomogeneous medium, 118
Interaction, 131, 233, 239, 243, 253, 259, 309
Interaction potential, 100
Interaction vertices, 124
Interference, 3–5, 11, 34, 106
Interference pattern, 173
Internal degree of freedom, 95, 179
Internal lines, 124
Intrinsic decoherence, 172
Isospin, 254
Isospin current, 240, 258
Isotope, 316

J

J/Psi, 220
 James Cronin, 278
 Jet events, 304
 Johannes Kepler, 185

K

Kaon, 218
 Kaon decay, 278
 Kepler orbits, 185
 Kinetic energy, 41, 144

L

Lagrangian, 46
 Lambda, 218
 Lamb shifts, 157
 Laplace—Runge—Lenz vector, 185
 Large Electron Positron collider, 305
 Large Hadron Collider, 222, 232, 255
 Laser, 196
 Lederman, 219
 Left-handed, 188, 204, 206, 208, 224, 231, 255, 264, 274, 277
 Length contraction, 66
 Lepton, 221
 Lifespan, 226
 Lifetime, 121, 129, 283
 Light cone, 64
 Llewellyn Thomas, 186
 Local symmetry, 99, 235, 321
 Longitudinal wave, 7
 Loop quantum gravity, 329
 Lorentz, 49
 Lorentz group, 187
 Lorentz invariance, 240, 281
 Lorentz invariant, 128, 254, 312
 Lorentz symmetric, 128, 188, 240
 Lorentz transformation, 128, 185, 321

M

Magnetic field, 101, 143, 179
 Magnetic moment, 189, 317
 Magnetic Resonance Imaging (MRI), 133
 Mass, 50, 114, 155, 205, 241, 256, 274, 282, 327
 Mass eigenstate, 272
 Massive particles, 55
 Massive waves, 50
 Massless particles, 55
 Mass potential, 92
 Mass shell, 56, 122

Matter, 149, 194, 279, 328

Max Planck, 3
 Maxwell, 3
 Measurement, 12, 164
 Medium, 10, 19, 44, 52, 61, 116, 253
 Mesons, 210, 220, 296, 307, 314
 Metaphors, 2
 Metric, 63
 Mexican hat potential, 242, 251
 Michelson and Morley, 25
 Minkowski, 63, 186
 Minkowski metric, 64
 Minkowski space, 64
 Mixing, 269
 Mixing amplitude, 271
 Mixing angle, 327
 Mixing fields, 248
 Mixing ratio, 254
 Möbius strip, 188, 196
 Modes, 86
 Momentum, 6, 114
 Momentum flux, 82
 Momentum space, 33
 Multiplets, 308
 Muon, 136, 218, 262

N

Near-field communication, 133
 Neddermeyer, 218
 Neutral current, 258, 264
 Neutrino, 219, 231, 261, 271, 276
 Neutrino mass, 273
 Neutrino oscillations, 271
 Neutrons, 21, 218, 266, 288, 315
 Newton, 3
 Nicola Cabibbo, 271
 Nitrogen, 267
 Noether current, 84, 213, 233
 Noether theorem, 84, 213
 Non-Abelian, 237, 299
 Nonet, 307
 Non-zero field, 253
 Nuclear force, 314
 Nuclear reactor, 219
 Nucleus, 21, 315

O

Observer, 124
 Octet, 295, 308
 Off mass shell, 122, 135
 Off-shell, 56, 170

Omega, 219
On mass shell, 122, 129, 135, 171
On-shell, 56
Orbital momentum, 180
Orthogonal, 89

P

Pair production, 125, 259
Parity, 204
Parity operation, 203
Parity violation, 276
Particle families, 282
Particle number conservation, 213
Particles, 3
Particle zoo, 217, 287
Path, 107
Path integral, 106, 140, 144, 326
Pentaquark, 225
Period, 36
Periodic boundary condition, 86
Perl, 220
PETRA, 305
Phase, 106, 144
Phase difference, 108
Phase shift, 95, 99, 119, 144
Phase velocity, 39
Phonon, 77, 87
Phonon density of states, 87
Photoelectric effect, 3
Photon, 5, 101, 104, 135, 218, 259
Pions, 218, 262, 288, 314, 318
Planck length, 324
Planck scale, 154
Planck's constant, 324
Plano, 219
Plasma, 61, 247
Poincaré group, 187
2-point propagator, 129
4-point propagator, 129
Pole, 129
Position space, 33
Positrons, 125, 149, 218
Potential, 41, 46, 51, 72, 74, 79, 89, 100, 110, 126, 138, 144, 232, 235, 239, 241, 250, 251, 253, 274, 292, 300, 324
Potential barrier, 170
Potential density, 79
Potential energy, 41, 59, 144
Powell, 218
Precession, 205
Pressure wave, 7

Probability, 30
Probability amplitude, 15
Probability distribution, 15
Product, 238, 289
Propagator, 129, 135
Protons, 21, 218, 266, 288, 313, 314, 317
Prowse, 219

Q

QED equation, 153
Qluon, 98
Quantization, 71, 85, 145, 180, 327
Quantization hypothesis, 71
Quantum, 71
Quantum algorithm, 175
Quantum chromodynamics, 287
Quantum computer, 173
Quantum decoherence, 34, 90, 161, 164, 177
Quantum Electrodynamics (QED), 101, 103
Quantum field theory, 17
Quantum loop gravity, 159
Quantum of space-time, 326
Quantum states of a particle that are indistinguishable, 14
Quark, 21, 98, 221, 287, 307, 309, 317
Quark–gluon interactions, 314
Quark–gluon plasma, 306
Quark jet, 303
Quark matter, 306
Qubits, 173

R

Radioactive decay, 267
Range, 257
Real particles, 37, 120, 134, 135
Real quantum, 129
Refractive index, 44, 116
Relativity, 124
Renormalisation group, 158
Renormalizable, 325
Renormalization, 154
Repulsive, 300
Repulsive force, 113, 140
Residual strong force, 314
Resonances, 120, 129, 134, 226, 287
Richard Feynman, 124
Richter, 220
Right-handed, 188, 204, 206, 208, 231, 255, 264, 274, 277
Right-hand rule, 201
Rochester, 218

Rolf Hagedorn, 306

Rotation, 187

Rutherford, 218

S

Samios, 220

Scalar, 20

Scalar field, 19, 240

Scalar particle, 256

Scattering, 220

Schrödinger, 161

Schrödinger's cat, 161, 178

Schwartz, 219

Screening, 156

Second quantization, 73, 145

Segre, 219

Selection problem, 176

Self-energy, 130

Self-interaction, 130, 265

Shear stress, 82

Sigma, 219

Singlet, 296

Solar neutrino deficit, 272

Source, 135

Space, 64

Space-like, 64

Space-time, 23, 187

Space-time translations, 187

Special relativity, 49, 185

Speed of light, 324

Spin, 95, 179, 185

Spin down, 188, 208

Spinor fields, 200

Spinors, 188, 200

Spin up, 188, 208

Spooky action at a distance, 164

Springs, 50, 73, 253

Standard model, 222

State, 95, 183, 192

Statistical sum, 140

Steinberger, 219

Strange, 220

Strangeness, 307

Strange quark, 278

String, 301

String theory, 159, 329

Strong force, 222, 276, 287, 317

Structure of the vacuum, 324

SU(2), 96, 225, 235, 282, 289

SU(3), 97, 225, 289, 299, 307

Sudbury Neutrino Observatory, 272

Super cooled, 248

Superfluidity, 210

Superposition, 10, 161, 173

Symmetric, 193, 289

Symmetric mix, 253

Symmetric wave function, 211

Symmetry, 84, 93, 95, 229, 232, 235, 247, 276, 289, 317, 321

Symmetry breaking, 231, 247, 258

Symmetry group, 93, 187

Symmetry operation, 93, 295

T

Tau, 137, 220

Tensor, 20

Tetraquark, 225

Tevatron, 313

Theory of relativity, 24

Thermodynamics, 148

Thomas rotation, 186

Thomson, 218

Time, 64

Time dilation, 66

Time direction, 125, 126, 148, 280

Time-like, 64

Time symmetry, 280

Ting, 220

Top, 221

Top quark, 275

Transformer, 133

Transition from quantum behaviour to classical behaviour, 166

Transmutation, 269

Transverse wave, 7

Travel back in time, 126, 148

Travelling salesman problem, 177

Tunnelling, 72, 170

U

U(1), 93, 98, 224, 232

Uncertainty principle, 146

Uncertainty relations, 31

Unitary rotation, 93

Up, 220

Up quark, 231, 261, 266, 288

Upsilon, 220

V

Vacuum, 19, 21, 43, 189, 249, 323

Vacuum diagram, 146

Vacuum polarisation, 156

Vacuum polarization, 136, 189, 260

- Val Fitch, 278
van der Waals force, 147
Vector, 20
Vector field, 20, 240
Velocity, 8, 50
Velocity of particles, 38
Vertex, 259
Virtual cloud, 189
Virtual electron, 119
Virtual gluon, 310
Virtual particle cloud, 137, 158, 163, 167, 170
Virtual particles, 37, 129, 135, 138
Virtual photon, 119, 129, 138, 153
- Waves, 3, 6, 135
W-bosons, 235, 261
Weak force, 217, 255, 257
Weak hypercharge, 232
Weak interaction, 276
Weak isospin, 235
Weak mixing angle, 271
Winding number, 104
Work, 56

X

- Xi, 219
X-rays, 218

Y

- Yukawa potential, 317

Z

- Z, 256, 265
Z-boson, 253
Zweig, 220

W

- W, 220
 $W^{+/-}$, 256
 W^+/W^- , 265
Wave function, 30, 165
Wavelength, 6
Wavelengths in water, 10
Wave packet, 30