Alois Lechicki (Fürth, Germany; E-mail: mail@lechicki.de)
Version 3.6
[Last revised: 15/01/2025]

If the Yang-Mills theory is the answer, what is the question?

Yang-Mills theory as understood by a layman

Summary

Yang-Mills theories play a central role in modern physics, forming the basis for the Standard Model (SM) of elementary particles and forces. The SM has proven to be an exceptionally successful theory, representing our most sophisticated theoretical framework for understanding the properties of matter and interactions at the fundamental level. In technical terms, interactions in the Standard Model are described by a Yang-Mills theory with the local $\mathrm{SU}(3)_c \times \mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ gauge symmetry. The objective of this article is to explain to non-physicists, who wish to learn something about contemporary physics, what this statement means. As the author is not a physicist, this text is intended as an introduction to Yang-Mills theory for a non-expert audience. However, given the sophistication of Yang-Mills theory, the paper assumes a fair bit of mathematical background. It includes some technical details, although these are largely presented in a superficial manner.

Contents

"... *I think I can safely say that nobody understands quantum mechanics. So do not take the lecture too seriously, feeling that you really have to understand in terms of some model what I am going to describe, but just relax and enjoy it. I am going to tell you what nature behaves like. If you will simply admit that maybe she does behave like this, you will find her a delightful, entrancing thing. Do not keep saying to yourself, if you can possibly avoid it, "But how can it be like that?" because you will get 'down the drain', into a blind alley from which nobody has escaped. Nobody knows how it can be like that.*" ([F9]) – Richard P. Feynman ([1])

1. Introduction

When reading popular science books or articles on modern physics, one often encounters the notion of *Yang-Mills theory*, which in turn has something to do with *gauge theory*. The Yang-Mills theory (YM) is actually a class of theories based on *gauge symmetry*. Yang-Mills theory appears to occupy a foundational position in quantum physics, providing the basis for the Standard Model (SM) of particle physics.

Prior to the formulation of gauge symmetry and Yang-Mills theory, numerous scholars, including Lorentz, Einstein, and Poincaré, had engaged in the study of symmetries within Maxwell's equations. They discovered a symmetry, the Lorentz symmetry, which is fundamental to the theory of special relativity (SR). This prompted other scientists to investigate whether there were additional symmetries inherent to Maxwell's equations. Hermann Weyl [W6] identified a novel symmetry of electromagnetism, now known as gauge symmetry ([N2]).

A little earlier, Einstein developed his general theory of relativity (GR). One of the key ideas of general relativity is the symmetry principle that the field equations should be invariant with respect to the choice of the (local) coordinate system. This is known *as the principle of general covariance*. From a modern perspective, this can be considered an example of gauge symmetry ([N2]).

In 1954, a seminal contribution to theoretical physics was made by Chen-Ning Yang and Robert L. Mills, who developed what is now known as Yang–Mills gauge theory through a creative generalisation of Maxwell's theory. However, for a period of almost twenty years, it remained in a state of dormancy as a beautiful but useless mathematical exercise. This situation changed in the 1970s, when, following significant advancements in both experimental and theoretical particle physics, it was called upon to unify the electromagnetic and weak interactions. Since that time, one of the most important guiding principles in physics has been that our description of the world should be based on a special type of classical field theory, namely Yang-Mills theory. With the exception of gravitation, all important theories of contemporary physics are quantum field theories (QFT) which in turn are quantised versions of Yang-Mills theories. Consequently, Yang-Mills theory now serves as the foundational framework for the Standard Model of elementary particles. ([H13], [N2])

Descriptions of Yang-Mills theories in the literature can be classified into two categories: those that are superficial and qualitative, and those that are mathematically precise, highly technical, and abstract.

- An example of the first category is:

Yang-Mills theory is basically a generalization of the principles that we use to understand the electromagnetic force. Because these principles were so successful with electromagnetism, it seemed natural to try to apply them to the other forces - the weak and strong nuclear forces. The Yang-Mills theory is, specifically, what is known as a gauge theory (...). In this theory space-time

[1] Richard Phillips Feynman (1918 –1988) was an American theoretical physicist. He received the Nobel Prize in Physics in 1965 jointly with Julian Schwinger and Shin'ichirō Tomonaga.
Julian Seymour Schwinger (1918 – 1994) was an American theoretical physicist.
Shin'ichirō Tomonaga (1906 – 1979), usually cited as Sin-Itiro Tomonaga in English, was a Japanese physicist.

fields have an internal symmetry: they are acted on by space-time dependant transformations in a way that leaves physical quantities, such as the action, invariant. These transformations are known as local gauge transformations. ([2])

While this reads quite well, it is not immediately clear what it actually means.

- An example for the second category is:

Yang-Mills theory is a gauge theory on a given 4-dimensional (pseudo-)Riemannian manifold X whose field is the Yang-Mills field – a cocycle $\nabla \in \mathbf{H}(X, \mathbf{B}\mathrm{U}(n))$ in differential nonabelian cohomology represented by a vector bundle with connection – and whose action functional is

$$\nabla \to \frac{1}{g^2}\int_X tr(F_\nabla \wedge * F_\nabla) + i\theta \int_X tr(F_\nabla \wedge F_\nabla)$$

for

- F_∇ *the field strength, locally the curvature* $\mathfrak{u}(n)$*-Lie algebra valued differential form on* X
- $\star$ *the Hodge star operator of the metric* g
- $1/g^2$ *the Yang-Mills coupling constant and* θ *the theta angle, some real number.* ([3])

Even if one is able to understand all of these mathematical terms, it is still not clear what the physical meaning of the whole thing is. What is the relationship between these concepts and the real world?

And you cannot expect too much help from Wikipedia either:

Yang-Mills theories are special examples of gauge theories with a non-abelian symmetry group given by the Lagrangian

$$\mathcal{L}_{gf} = -\frac{1}{2}\mathrm{Tr}(F^2) = -\frac{1}{4}F^{a\mu\nu}F^a_{\mu\nu}$$

with the generators T^a of *the Lie algebra, indexed by a, corresponding to the F-quantities (the curvature or field-strength form) satisfying*

$$\mathrm{Tr}(T^a T^b) = \frac{1}{2}\delta^{ab}, [T^a, T^b] = if^{abc}T^c.$$ ([4])

But again, it does not mean much unless one already knew the answer anyway.

In this paper, I will try to fill the gap between general, qualitative on the one hand and mathematically abstract on the other hand expositions of the Yang-Mills theory. Unfortunately, the Yang-Mills theory is a complex mathematical theory. So just to get a superficial understanding of the theory it is essential to grasp its mathematical formalism, at least to a certain extent ([5]). The language of physics is mathematics, and this fact cannot be escaped, even in a 'layman exposition' such as this article. That does not mean, however, that the reader needs to comprehend all the equations. It is possible to understand the essence of the discussion without having a detailed knowledge of the equations.

---

[2] https://www.reddit.com/r/askscience/comments/2u6jgu/can_you_simplify_and_explain_whats_behind

[3] https://ncatlab.org/nlab/show/Yang-Mills+theory

[4] If you wonder why this description is apparently quite different from the previous one, remember what Feynman said: "*every good physicists know five different mathematics to describe the same phenomenon.*"

[5] It is important to remember that quantum physics is fundamentally non-intuitive. So it is one of those cases where mathematics is essential. At the same time, mathematics is the main problem in understanding advanced physics. The precise description of physical phenomena requires a mathematical arsenal that is not available to everyone.
However, physics is not mathematics. Besides being mathematically consistent a successful physical theory must also be consistent with known data, which is a stringent demand. "*It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong.*" -- Richard P. Feynman.
Another thing is the physical intuition that is required to understand how the Universe works.

My description of the Yang-Mills theory is from a mathematical viewpoint largely cursory and does not delve deeply into the intricacies of the theory. However, from time to time it will be necessary to go into the 'engine room' of the theory. Although one can see a lot of very interesting things there, we will not spend an extensive amount of time in this area.

It should be noted that the explanations within this text do not claim to be mathematically precise. The presentation of facts has been simplified to a considerable degree, and the accuracy of the information provided may be limited due to the author's lack of expertise in the subject matter.

The material in this paper is not original; some comes from primary literature, but the majority of what I write is taken from the existing textbooks. The textbooks [A1], [F2–F6], [S2–S3] and [S16] were especially influential.

Finally, it is important to note that any copyright issues or inaccurate attribution are unintentional. Should any such concerns arise, or if there are comments to be shared, they can be directed to the email address provided on the first page of this article.

2. Prelude on the Standard Model (SM)

The question of the fundamental nature of reality has been a topic of interest for humans for a considerable length of time. What are the fundamental constituents of the universe? And what are the underlying principles that bind them together? Why is it that so many things in this world exhibit similar characteristics? It has been established that the matter of the world is constituted by a small number of fundamental building blocks of nature. The universe exists as a consequence of the interactions between these fundamental particles. These interactions encompass attractive and repulsive forces, decay, and annihilation. All forces in the universe can be attributed to four fundamental interactions between particles. ([C6])

In the 1970s physicists developed a theory of fundamental particles (also called elementary particles) and their interactions (except gravity) called *The Standard Model* (SM). In a nutshell, the picture is as follows. The fundamental matter units are *fermions* which are structureless at the smallest distances currently probed by the highest-energy accelerators. Fermions interact via the exchange of gauge field quanta. All the non-gravitational interactions we know of are described by Yang-Mills gauge theories. Consequently, these Yang-Mills theories provide an answer to the question *What holds the world together?* The main goal of this paper is to discuss the particular nature of these Yang-Mills theories. Gauge theories, however, have a very rich mathematical structure, at the classical and especially at the quantum level. Within the scope of this article, we can only limit ourselves to just a few elementary aspects.

In this chapter, we look briefly at the Standard Model (see [A1] and [W0] as a general reference for this chapter).

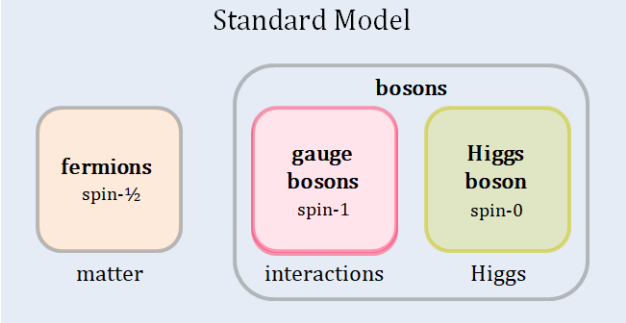
The Standard Model (SM) is a theory — or rather a set of theories — that encapsulates and explains all the experimental studies of the microworld that have ever been made. It is the most complete explanation of the fundamental particles and interactions to date. All the phenomena of Maxwell's electromagnetism, all the studies of radioactivity, all the data produced in particle accelerator laboratories — all of this can be explained by the Standard Model ([6]). It contains just two basic components.

First, there are some fundamental particles — the quanta of the fundamental fields. There are two types of fundamental fields: *fermions* and *bosons* ([7]). Bosons are split into two groups: *gauge* (or *force-carrying*) *bosons* and *Higgs boson*. In a sense, the fermions are *particles of*

[6] Recently, however, experiments have hinted at effects that may lie beyond the Standard Model — see Remark 2.4.

[7] This nomenclature was coined by Paul A. Dirac, who named it in honour of two renowned scientists: Enrico Fermi (1901 – 1954) and Satyendranath Bose (1894 - 1974).

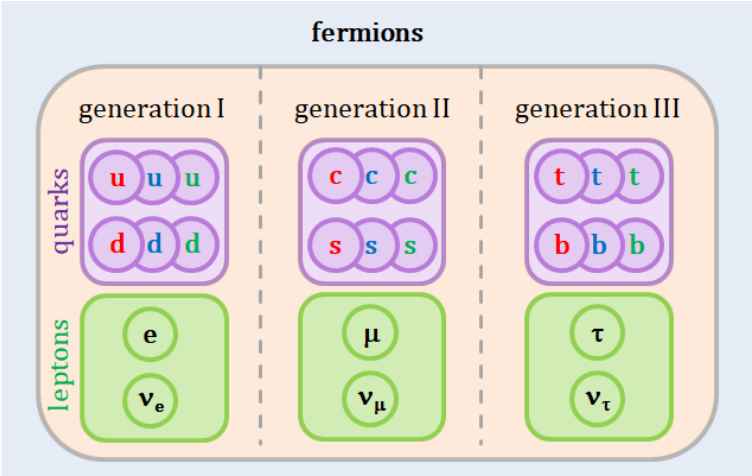
matter while the gauge bosons are *particles of interactions*. The difference between the Higgs boson and the gauge bosons is that the latter are associated with gauge symmetries, while the Higgs is not. The Higgs field provides a mechanism for charged fermions and weak gauge bosons to acquire nonzero masses (see Section 7.4). In this regard, the Higgs field is not considered to form a fundamental force.



The fermions, with spin-½ (in units of $\hbar$ – see Section 3.7), are of two types: three generations (families) of *lepton doublets* (the electron e and its neutrino ν_e, the muon μ and its neutrino ν_μ, and the tau τ and its neutrino ν_τ) ([8]) and three generations (families) of *quark doublets*

I. the *up* (u) and *down* (d) quarks,
II. the *charm* (c) and *strange* (s) quarks, and
III. the *top* (t) and *bottom* (b) quarks.

Each quark comes in three varieties, distinguished by *colour*. It is precisely this quantum number that underlies the dynamics of the strong interactions. Colour, in fact, is a kind of generalized charge, for the strong interactions ([9]). One denotes the three colours of a quark by 'red', 'blue', and 'green'. Thus we have the triplet (u, u, u), and similarly for all the other quarks (see Section 7.5).



Each fermion has an antiparticle (denoted, e.g. $\bar{u}$, $\bar{d}$, etc.). Moreover, each fermion except neutrinos exists in a left-chiral state as well as in a right-chiral state, which corresponds to a different behaviour related to the weak interaction (see Section 5.7).

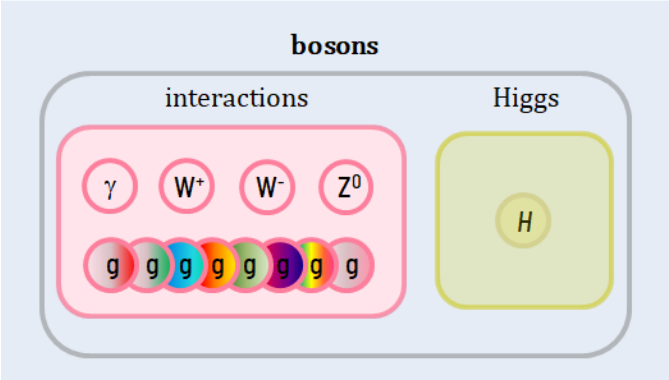
All stable matter in the universe is made from particles that belong to the first generation; any heavier particles quickly decay to the next most stable level.

---

[8] The historical nomenclature of leptons was derived from their perceived lightness; however, this is no longer true with the discovery of the tau, which is about twice as heavy as the proton ([H14]).

[9] For the nuclear forces physicists use the term 'interaction(s)' both in the singular and in the plural, whereas for gravitation and electromagnetism only the singular is common. This may have to do with the fact that in the Standard Model there are several quanta for the strong and weak nuclear, while there is only one photon and (probably) only one graviton. ([E1])

The gauge bosons, all of which possess spin 1, are the massless photon (γ), the three massive weak bosons (W^+, W^-, and Z^0), and the eight massless gluons (**g**). Gluons are electrically neutral, but they are not colour neutral. Each gluon carries one colour and an anticolour (see Section 7.5). The massive Higgs boson (H) corresponds to a scalar field so it has spin 0 (see Section 7.4). It is electrically neutral. The bosons serve as their own antiparticles.



From these few (eighteen, if we ignore colour of quarks and gluons ([10])) basic building blocks, all the familiar matter in the Universe can be constructed ([11]).

Second, these fundamental particles interact in only three ways ([12]):

- via the *electromagnetic interaction* which, as we shall see later (Section 7.4), is an aspect of the unified *electroweak interaction*. Electromagnetism acts on any particle that carries electric charge — the carrier particle is *photon*.
- via the *weak interaction*, another aspect of the unified electroweak interaction. The weak interaction acts on all particles (it is the only interaction felt by the neutrinos) but its effects are often masked by the other interactions; the carrier particles here are the W^+, W^- and Z^0 bosons.
- via the *strong interaction*. The strong interaction acts on any particle that carries *colour charge*. In other words, it acts only on the quarks and gluons. Quarks and gluons do not appear as free particles. They form a large number of bound states, the *hadrons*. It is the strong force that binds quarks into protons, neutrons, mesons and the like. The carrier particle in this case is the *gluon* (see Section 7.5).

Each of the carrier particles couples to the charges ([13]) with a characteristic 'stickiness' known as the *coupling constant* ([14]) for the interaction. The coupling constants are different, so the strengths of the three interactions are different.

The hadrons are of two types: hadrons with spins $\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots$ (i.e. fermions) are *baryons*, those with spins 0, 1, 2, . . . (i.e. bosons) are *mesons*. Examples of baryons are *nucleons* — the neutron n and the proton p. Baryons contain three quarks, while mesons are quark-antiquark systems. One

[10] Of course, there are still antiparticles to be taken into account.

[11] It turns out that roughly 68% of the universe is *dark energy*. *Dark matter* makes up about 27%. The rest – everything on Earth, everything ever observed with all of our instruments, all normal matter – adds up to less than 5% of the universe. Neither dark matter nor dark energy is included in the Standard Model.

[12] Gravitational interaction is not part of the SM.

[13] To every force, there is a corresponding *charge*. For the electromagnetic force, that charge is the well-known *electric charge*. The charge for the strong force is the *colour charge*, e.g. quarks have this charge. The case of the weak force is more complicated. The problem with 'weak charges' is that electroweak symmetry is spontaneously broken – see Section 7.4. The corresponding charge for gravitation is the energy-content of a system.

[14] Coupling constants are not, in fact, constant. The actual value of a coupling constant changes if one changes the reference point, i.e. the energy scale. Nevertheless, the names coupling constant, fine-structure constant, etc. are still being used.

immediate consequence is that quarks have fractional electromagnetic charge. For example, the proton has two u quarks of charge $+\frac{2}{3}$ and one d quark of charge $-\frac{1}{3}$. The neutron has the combination (ddu), while the meson π^+ has one u and one anti-d, i.e. ($u\bar{d}$), and so on.

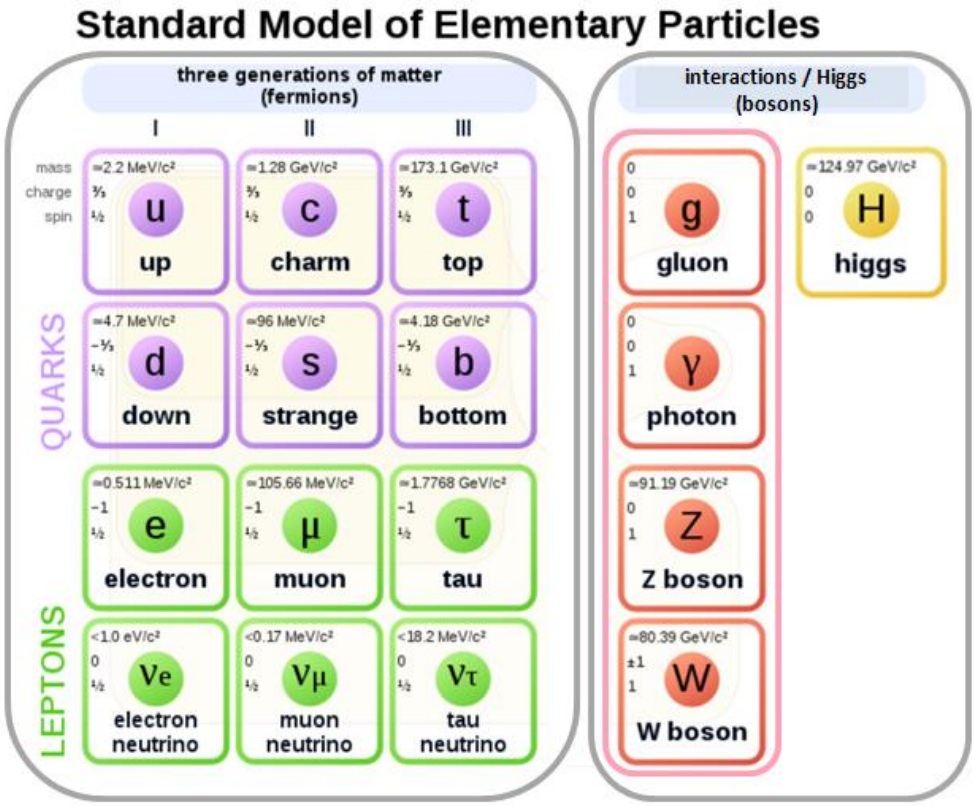
Unlike the constituents of atoms and nuclei, quarks have not been observed as stable isolated particles. When hadrons of the highest energies currently available are smashed into each other, what is observed downstream is only lots more hadrons, not fractionally charged quarks. The explanation for this novel behaviour of quarks is now believed to lie in the nature of the strong force – see Section 7.5.

Another property of hadrons is that they have no net colour charge (i.e. they are colour-neutral) even though the quarks themselves carry colour charge. So what holds the nucleus together when positive protons repel each other with electromagnetic force, and protons and neutrons are colour-neutral? The answer is, in short, that the strong force between the quarks in one proton and the quarks in another proton is strong enough to overwhelm the repulsive electromagnetic force. This is called the *residual strong interaction*, and it is what 'glues' the nucleus together.

The behaviour of the fundamental interactions is described completely by relativistic quantum field theories of the Yang-Mills type. Quantum field theory – or QFT for short – is the fundamental formal and conceptual framework of the Standard Model. The weak and electromagnetic interactions of both quarks and leptons are described in a (partially) unified way by the *electroweak theory* of Glashow, Salam and Weinberg (GSW), which is a generalization of *quantum electrodynamics* (QED). The strong interactions of quarks are described by *quantum chromodynamics* (QCD), which is also analogous to QED. The similarity with QED lies in the fact that all three interactions are types of gauge theories, though realized in different ways.

Thus, at a deep level, there are only two interactions that are relevant for fundamental particles: the electroweak and strong interactions.

A summary of fundamental particles as described by the Standard Model (we ignore colour of quarks and gluons) is given below (from Wikimedia Commons):



A summary of interactions is given in the table below (the gravitational force is added for completeness but is not part of the Standard Model).

Interactions		Relative strength	Exchange particle (gauge boson)	Particles acted upon	Range
strong		1	eight gluons	quarks & gluons	10^{-15} m
electroweak	electro-magnetic	1/137	photon	electricaly charged particles	∞
	weak	10^{-10}	W^+, W^- and Z^0 bosons	fermions (i.e. leptons & quarks)	10^{-18} m
gravitational		10^{-38}	graviton (hypothetical)	all particles	∞

Remark 2.1. The notion of *spin* requires some comment: fundamental particles have an intrinsic spin angular momentum ([15]). The adjective intrinsic means that they do not have spin because someone is spinning them. They just spin – or rather, they just have a measurable quantity with the same units as angular momentum. Spin is an internal degree of freedom of a particle. In current physics, fundamental particles are featureless – like a mathematical point. In order for something to be perceived as spinning, the thing spinning would need something like a 'front' and a 'back'. Featureless, point particles do not have anything like that. Spin is a convenient label for a measurable quantity and not a description of reality ([16]) ([P6]). Every fundamental particle has associated with it a *spin quantum number* S (often called the *spin number* or just the *spin*), where S is any whole number multiple of a half (in units of $\hbar$). Fermions have half integral spin ($\frac{1}{2}$, $\frac{3}{2}$, $\frac{5}{2}$, etc.) and bosons have integral spin numbers (0, 1, 2, etc.). No spin numbers are possible in between these. Spin is a quantised quantity.

An important fact about spin is that it cannot be changed. The spin of a particle is a fundamental unchangeable property of it just like its mass or electric charge. Particles made from combinations of fundamental ones will have an overall spin that is a combination of the individual spins.

The fundamental fermions have a spin of ½. Fermions include (in addition to leptons and baryons) nuclei of odd mass number (e.g., tritium, helium-3, uranium-233, etc.). For reasons we do not fully understand, a consequence of the odd half-integer spin is that fermions obey the *Pauli Exclusion Principle,* which forbids more than one particle of this type from occupying a single quantum state (see Section 5.8). Therefore, fermions cannot co-exist in the same state at the same location at the same time. The basic rule is *'don't sit where I'm sitting'*. This rule underlies, for example, the build-up of electrons within an atom in successive orbitals around the nucleus and thereby prevents matter from collapsing to an extremely dense state. Consequently, the exclusion principle allows fermions to build everything from atoms to planets.

The force carrying bosons (gluons, photons, and the $W^{\pm}$ and Z^0 bosons) have spin 1 since they go with vector fields. The Higgs boson corresponds to a scalar field so it has spin 0. If the particle of the gravitational field is ever discovered, it would be called a *graviton* and would have spin 2 since it corresponds to a tensor field ([F8]). In addition to these force carrying particles, bosons include mesons (e.g., pions and kaons) and nuclei of even mass number (e.g., helium-4). Bosons differ significantly from fermions in that there is no limit to the number that can occupy the same quantum state (they obey the *Bose-Einstein statistics,* whereas fermions obey the *Fermi-Dirac statistics*). This behaviour gives rise, for example, to the remarkable properties of helium-4 when it is cooled to become a superfluid.

---

[15] *Angular momentum* is the rotational equivalent of linear momentum. It is an important quantity in physics because it is a conserved quantity – the total angular momentum of a closed system remains constant. In three dimensions, the angular momentum for a particle is the cross product $\boldsymbol{x}\times\boldsymbol{p}$ of the particle's position vector $\boldsymbol{x}$ (relative to some origin) and its momentum vector $\boldsymbol{p}$ ([W11]).

[16] It may be tempting to think about fundamental particles as little spinning balls, but such a picture quickly leads to paradoxical results. One of the surprises of modern science is that atoms and sub-atomic particles do not behave like anything we see in the everyday world. Matter is not composed of particles in the classical sense, nor is it an ordinary wave. These terms are used metaphorically and should not be interpreted literally.

The spin of a particle determines which mathematical tool we need to describe its field: we describe spin-0 particles using scalars, spin-½ particles using spinors ([17]) and spin-1 particles using vectors.

Remark 2.2. Each of leptons comes in three generations, distinguished by a quantum number called *lepton flavour*. Thus we have electron flavour L_e, muon flavour L_μ and tau flavour L_τ. Each is postulated to be conserved in all leptonic processes. The electromagnetic interactions of the μ and the τ leptons are the same as for the electron e. In weak interactions, each lepton (e, μ, τ) is accompanied by its 'own' neutral partner, a *neutrino*. In the Standard Model, the three neutrinos are assigned lepton flavour quantum numbers in such a way as to conserve each lepton flavour separately.

Similarly like the leptons, also quarks (u, d, s, c, b, t) carry flavour quantum numbers called *quark flavour*. Thus quarks have both 'colour' and 'flavour' ([18]). The strong and electromagnetic interactions of quarks are independent of quark flavour and depend only on the electromagnetic charge and the strong charge (i.e. colour), respectively. This means, in particular, that flavour cannot change in a strong interaction among hadrons – that is, flavour is conserved in such interactions. In weak interactions, by contrast, quark flavour is generally not conserved.

Remark 2.3. The Standard Model of fundamental particles is actually a misnomer ([H7], [T5]). This is because it is, in fact, the Standard Model of quantum field physics. The fundamental objects of the Standard Model are all fields ([19]). Take for example the quantum electrodynamics (QED): this theory has two fundamental objects, the electromagnetic field A_μ and the electron-positron (Dirac) field ψ (see Section 5.8). In general, the fundamental entity in Quantum Field Theory (QFT) is not the particle, but rather the field. The field is the property of spacetime that in the presence of energy and momentum a particle can be created. Particles are the quantised excitations of the underlying fields (the field quanta). It means that we associate these excitations with what we perceive as particles ([20]). So remember, any time we are talking about particles, we are in fact talking about the idealised excitation of the field that goes with that particle. Interactions between fields create or destroy such excitations. So for instance, a photon may be created when the field of electrons interacts with the electromagnetic field.

The different particle types are truly separated in QFT: each type is represented by one field, and the fields interact. These interactions are quantified by the Lagrangian density, which

---

[17] Spinors are discussed in Section 5.7.

[18] The term 'flavour' was coined in 1971 by Murray Gell-Mann and his student Harald Fritzsch at a Baskin-Robbins ice-cream store in Pasadena, for describing the different types of quarks. Just as ice cream has both colour and flavour so do quarks ([K5]).
Fritzsch had been born in Zwickau, south of Leipzig in East Germany. Together with a colleague he had defected from Communist East Germany, escaping from the authorities in Bulgaria in a kayak fitted with an outboard motor. They had travelled 200 miles down the Black Sea to Turkey ([B1]).

[19] For a detailed discussion of this issue, please refer to [H7].

[20] The idea of a particle as an excitation of a quantum field is not particularly intuitive (not to mention even more abstract definition: a particle is an *irreducible representation of a symmetry group*). R. Feynman recollects ([F11]):
"I went to MIT. I went to Princeton. I came home, and he [father] said, '*Now you've got a science education. I have always wanted to know something that I have never understood, and so, my son, I want you to explain it to me*.'
I said yes.
He said, '*I understand that they say that light is emitted from an atom when it goes from one state to another, from an excited state to a state of lower energy*.'
I said, '*That's right*.'
'*And light is a kind of particle, a photon, I think they call it*.'
'*Yes*.'
'*So if the photon comes out of the atom when it goes from the excited to the lower state, the photon must have been in the atom in the excited state*.'
I said, '*Well, no*.'
He said, '*Well, how do you look at it so you can think of a particle photon coming out without it having been in there in the excited state* ?'
I thought a few minutes, and I said, '*I'm sorry; I don't know. I can't explain it to you*.'
He was very disappointed after all these years and years of trying to teach me something, that it came out with such poor results."

essentially determines everything about the theory (see Section 5.1). Fields are physical reality – particles are observer-dependent. For example, an accelerating observer may see field excitations, i.e. particles, where an inertial observer sees nothing. Saying, e.g. that 'particles collide' is to imply that the underlying fields interact through these excitations, whereby energy and momentum are exchanged, some excitations are annihilated and new excitations are created. All fields exist at all points in time and space. ([T5])

Take note that there are no macroscopic fields except for the electromagnetic field (and gravitation). The fundamental excitations of all spinor (i.e. matter) fields can never occupy the same state (due to the Pauli exclusion principle – see Remark 2.1) and thus cannot reinforce one another to produce a macroscopic field. The only known fundamental scalar field (the Higgs field) is massive and thus does not operate on macroscopic scales. And finally, all gauge fields (except for the electromagnetic field) also cannot operate at macroscopic scales. For the gauge field responsible for weak interactions, the reason is again that it is a massive field. For the strong interaction field, the reason is called 'confinement' as we shall see in Section 7.5. ([N1])

Remark 2.4. The Standard Model (SM) has been extremely successful. In fact, it explained almost all experimental results and precisely predicted a wide variety of phenomena. One exception is recent data indicating that neutrinos have mass, although the SM assumes that neutrinos are massless. Scientists have found that the three neutrinos oscillate, or transform into one another, as they move. This property is only possible because neutrinos are not massless after all. However, even this can be easily accommodated by an extension to the model.

Still, there are many questions for which the Standard Model has no answer. For example, why there are three generations of matter particles – this is the so-called *flavour problem*. In the everyday world, it seems that only the first-generation particles (i.e. the electrons and the up and down quarks) play crucial roles. In view of the simplicity of the fundamental laws of physics, the multiple generations of quarks and leptons seem unnecessary. This question, succinctly expressed by the famous quip of I.I. Rabi ([21]), *"who ordered that?"* ([22]), uttered in connection with the discovery of the muon, has been exacerbated by the discovery of the third generation. ([I1])

Much worse than the number of particles is the fact that the masses of all these particles have to be put into the Standard Model 'by hand'. In other words, physicists measure these values experimentally and manually plug the results into the equations. The Standard Model tells us nothing about what these masses should be. We know that it is the coupling of the fermions fields to the Higgs field that gives the various fermions their mass. But we cannot calculate the strength of these couplings. ([W0])

Furthermore, the SM says nothing about gravity. The problem is that when one tries to construct a theory of gravitation using the language of quantum field theory, the resulting theory is non-renormalisable. It means that it produces nonsensical infinities, which cannot be eliminated using known mathematical methods, which worked fine in the case of all other fields (see Section 7.3 and 7.5) ([23]).

And in the final instance, physicists now believe that about 95% of the universe is not made of ordinary matter as we know it. Instead, much of the universe consists of dark matter and dark energy that do not fit into the Standard Model.

---

[21] Isidor Isaac Rabi (1898 – 1988) was an American physicist who won the Nobel Prize in Physics in 1944 for his discovery of nuclear magnetic resonance.

[22] Willis Lamb echoed this sense of frustration in his 1955 Nobel lecture when he said: "*…the finder of a new elementary particle used to be rewarded by a Nobel Prize, but such a discovery now ought to be punished by a $10,000 fine*." ([B1]). Willis Eugene Lamb Jr. (1913 – 2008) was an American physicist who won the Nobel Prize in Physics in 1955.

[23] Feynman discovered that any direct quantisation of gravity leads to an infinite numbers of infinities rendering the theory non-renormalisable.

3. Basic concepts

In this chapter, I discuss the meaning of a few very important concepts which we shall use in the following chapters (see [A1], [S2] and [S17] as a general reference for this chapter).

3.1. Spacetime

In Newtonian physics, space and time never get mixed up. The two concepts are separate and distinct. To Newton, three-dimensional space was space, and time was universal time. They were entirely separate, the difference being absolute ([S17]). The Galilean transformation of classical mechanics expresses the assumption of a universal time independent of the relative motion of different observers. Thus the transformation guarantees that the time at which any event happens is the same in all inertial frames. In relativistic physics, however, space and time become intertwined through the Lorentz transformations (see Section 3.2). Time intervals and space intervals are not the same to all observers, but instead become mixed with one another. What is purely a distance to one observer may correspond to both a distance and a time interval to an observer in a different frame of reference. ([H4], [W11])

Spacetime is a conceptual model combining the three dimensions of space with the fourth dimension of time. To describe the location of an object, we introduce a coordinate system (i.e. a *reference frame*). Each location is then described by three numbers (x, y, z) and the distance ds between any two objects can be calculated using the Euclidean metric $ds^2 = \Delta x^2 + \Delta y^2 + \Delta z^2$ (where ds^2 stands for $(ds)^2$ and Δx^2 for $(\Delta x)^2$ and so on).

However, in physics, we are not only interested in knowing *where* something happens but also *when*. Therefore, to properly describe an object in physics, we actually need four numbers (t, x, y, z), where t is the time coordinate. This means that we consider a four-dimensional space with a *spacetime* coordinate system. Thus a reference frame is a coordinate system for both space and time. A reference frame is chosen as a matter of convenience, and generally varies from one observer to another. To relate data from different frames, we need a rule that translates the reading in one frame to another. This is called a *transformation law*.

While (x, y, z) describes the location of an object in space, (t, x, y, z) describes the location of an *event* in spacetime ([24]). Now the problem is that the differences in the spatial components (Δx, Δy, Δz) are measured e. g. in meters, while Δt is measured in time units. To fix this problem one introduces a constant c which has units of meter per second: (ct, x, y, z).

The constant c encodes the maximum speed at which anything can travel in spacetime. It is called the speed of light because light travels (in vacuum) at this maximum speed. The constant c is essential since time and space components can get mixed if we change the coordinate system. Note: ct is the distance light travels in time t.

Now, what is the distance between two events in spacetime? Using Pythagorean theorem we might write down (the symbol := means that we are dealing with a definition)

$$ds^2 := c^2\Delta t^2 + \Delta x^2 + \Delta y^2 + \Delta z^2. \tag{3.1}$$

But as Einstein figured out this formula is NOT the right one. Instead, the correct expression for the spacetime distance between events (called *spacetime interval*) is

$$ds^2 := c^2\Delta t^2 - \Delta x^2 - \Delta y^2 - \Delta z^2. \tag{3.2}$$

It means that the local structure of spacetime is not Euclidean (3.1) but Minkowskian (3.2). What makes Minkowski ([25]) space different from Euclidean space is the way we define distances

---

[24] An event is something that happens independently of the reference frame that might be used to describe it. There are many sets of coordinates and therefore many descriptions of the same events. The principle of relativity means that the laws governing those events are the same in all inertial reference frames. An inertial frame is one in which a particle, with no external forces acting on it, moves in a straight line with uniform velocity. ([S16])

[25] Hermann Minkowski (1864 – 1909) was a mathematician and professor at Königsberg, Zürich and Gottingen. In different sources Minkowski's nationality is variously given as German, Polish, Lithuanian-German or Russian. At the Eidgenössische Polytechnikum, today the ETH Zurich, he was one of Einstein's teachers ([W11]).

in it. We denote the 4-dimensional Euclidean space as $\mathbb{E}^4$ and the 4-dimensional Minkowski space as $\mathbb{E}^{1,3}$ ($= \mathbb{E}^1 \times \mathbb{E}^3$), or shortly as $\mathbb{M}$ ([26]).

Minus sign in the definition (3.2) of the Minkowski metric ([27]) is motivated by the physical meaning of ds^2. The spacetime interval ds^2 between two events is equal to the (squared) time interval measured by an observer for whom the object in question appears at rest (multiplied by the constant c^2) ([28]). And since in special relativity ds^2 must be independent of the inertial frame of reference ([29]) (i.e. unchanged when the coordinates are Lorentz transformed ([30])), we have to take minus sign when defining it ([S2]).

Example 3.1. ([H5]) Consider the following events:

Event A: a solar flare erupts on the sun.

Event B: an astronomer witnesses the flare from an observatory on Earth.

Event C: the astronomer takes a drink 5 minutes (300 seconds) after the solar flare occurred (i.e. 200 seconds before he sees the flare).

Let us first compute the spacetime interval ds^2 between A and B. Since light travels directly from event A to event B, these two events are light-like separated, so the spacetime interval must be zero ([31]). Indeed, the sun is approximately 150 million kilometres from Earth (as judged from our reference frame), and it takes about 500 seconds (8.3 minutes) for light to travel that distance, so the astronomer will see the flare 500 seconds after it occurred. The spacetime interval AB between events A and B is:

$$ds^2 = c^2\Delta t^2 - \Delta x^2 = (300{,}000\ \text{km/s})^2(500\ \text{s})^2 - (150{,}000{,}000\ \text{km})^2$$
$$= (150{,}000{,}000\ \text{km})^2 - (150{,}000{,}000\ \text{km})^2 = 0.$$

It is not quite intuitive that the spacetime interval between two events can be zero although their spatial distance is huge.

Now let us compute the spacetime interval between A and C. It takes 500 seconds for light to travel between the sun and Earth, so one would have to travel faster than light to get from the

---

[26] We think of time, as well as physical space, as being a 'Euclidean geometry', rather than as being just a copy of the real line $\mathbb{R}$. This is because both $\mathbb{R}$ and $\mathbb{R}^3$ have preferred origin, whereas in Euclidean or Minkowskian geometry there is no such preferred element (see e.g. [P5] Chap. 17).

[27] We say that the metric signature is $(+, -, -, -)$. Equivalently, one may define ds^2 with respect to the signature $(-, +, +, +)$. This is only a matter of convention. The choice of signature is given a variety of names, for example:

- $(+, -, -, -)$ - West Coast metric, Feynman's favourite
- $(-, +, +, +)$ - East Coast metric. The choice of Pauli, Weinberg and Schwinger.

Let me note that most particle physicists use the signature $(+, -, -, -)$ but most relativists use $(-, +, +, +)$, and you need to be careful when reading equations. Notice that the Minkowski metric ds^2 is not actually a metric from the viewpoint of topology, since it can take negative values. It is sometimes called pseudo-Riemannian metric. More precisely, to define the distance one has to get square root of ds^2, which leads to possibly imaginary distances. Hence terms norm and distance are usually avoided and replaced with *spacetime interval* ([W11]).

[28] This time is called *proper time* and denoted by τ. τ^2 is just the negative of ds^2, and is therefore an invariant ([S17]).

[29] Recall that it is a reference frame (i.e. a set of coordinates) in which the observers are not subject to any accelerating force. In special relativity, time measurements in inertial frames that are not at rest with respect to each other are not equivalent; each inertial frame must have its own time coordinate, the value of which is the time as read off a standard clock at rest in that frame (= proper time). This follows from the experimental fact that all (inertial) observers measure exactly the same value for the speed of light ([D1]).

[30] Lorentz transformations are discussed in Section 3.2.

[31] Events with zero spacetime separation, $ds^2 = 0$, are called *light-like* separated. Such events are causally related, and all observers will agree that they can be connected by a light ray.

flare to the event C. Therefore, events A and C are space-like separated ([32]), and the spacetime interval AC must be negative. Let us check:

$$ds^2 = c^2\Delta t^2 - \Delta x^2 = (300{,}000\ \text{km/s})^2(300\ \text{s})^2 - (150{,}000{,}000\ \text{km})^2$$
$$= -1.44 \times 10^{16}\ \text{km} < 0.$$

Finally, let us compute the spacetime interval between B and C. Since the events B and C occurred in the same location, these two events must be time-like separated ([33]). Let us verify that the spacetime interval BC is positive:

$$ds^2 = c^2\Delta t^2 - \Delta x^2 = (300{,}000\ \text{km/s})^2(200\ \text{s})^2 - 0^2 = 3.6 \times 10^{15}\ \text{km} > 0.$$

Let us consider the following 4×4 matrix

$$\eta^{\mu\nu} := \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix},$$

where by convention Greek indices μ, ν run from 0 to 3, so $\eta^{00} = 1$, $\eta^{11} = -1$, $\eta^{22} = -1$, $\eta^{21} = 0$, etc. Putting $\Delta x_0 = c\Delta t$, $\Delta x_1 = \Delta x$, $\Delta x_2 = \Delta y$ and $\Delta x_3 = \Delta z$ we can rewrite (3.2) as follows

$$ds^2 = \begin{bmatrix} \Delta x_0 & \Delta x_1 & \Delta x_2 & \Delta x_3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \Delta x_0 \\ \Delta x_1 \\ \Delta x_2 \\ \Delta x_3 \end{bmatrix}.$$

Thus the Minkowski metric (3.2) is conventionally denoted by $\eta^{\mu\nu}$ (and identified with this matrix). In addition, it is also conventional to denote (column) 4-vectors simply by a subscript Greek letter:

$$x_\mu = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} ct \\ x \\ y \\ z \end{bmatrix}.$$

This might be confusing because exactly the same symbol x_μ is used for the vector (or $\eta^{\mu\nu}$ for the matrix) and its components. The difference between vectors and components of vectors is not always clearly stated in literature and is to be deduced from the context.

In contrast, the usual 3-vectors, e.g. three-component vectors (x, y, z) that define the location of an object in space (not spacetime), are conventionally denoted by a little arrow on top of them $\vec{x}$ or by using italic bold letters ([34]): $\boldsymbol{x}$. Thus we will sometimes write $x_\mu = (ct, \boldsymbol{x})$. For the sake of simplicity, x will be used to denote x_μ when no confusion arises.

As we have already mentioned, the zeroth component x_0 of x_μ is in physics interpreted as the time component. This is an arbitrary choice and one could equally take any other component as the time coordinate.

Using this index notation we can write the Minkowski metric as inner product ([35]) in Minkowski space:

---

[32] Events with a negative spacetime separation, $ds^2 < 0$, are called *space-like* separated. Such events are not causally related.

[33] Events with a positive spacetime separation, $ds^2 > 0$, are called *time-like* separated. Such events are causally related. They are closer together in space than they are in time.

[34] Through the paper italic bold letters will denote 3-vectors. The components of a 3-vector are written as (x, y, z) or (x_i), or (x_1, x_2, x_3). In general the Roman indices $i, j, k, \ldots$, do not include the time component. Greek indices μ, ν, ... refer to spacetime and take values 0, 1, 2, 3 or t, x, y, z.

[35] Note that the inner Minkowski product is not an inner product in the usual sense, since it is not positive-definite. These misnomers, 'Minkowski metric' and 'Minkowski inner product', conflict with the standard meanings of metric and inner product in pure mathematics.
Instead of 'Minkowski inner product' we will also say 'scalar product' for short.

(3.3) $$ds^2 = \Delta x_\mu \eta^{\mu\nu} \Delta x_\nu := \sum_{\mu=0}^{3} \sum_{\nu=0}^{3} \eta^{\mu\nu} \Delta x_\mu \Delta x_\nu.$$

It is convenient to introduce superscript indices to avoid writing the Minkowski metric all the time:

$$x^\mu = [x^0 \quad x^1 \quad x^2 \quad x^3] := \eta^{\mu\nu} \Delta x_\nu = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} ct \\ x \\ y \\ z \end{bmatrix} = [ct \quad -x \quad -y \quad -z].$$

This is known as *raising an index* ($x_\mu \to x^\mu$). Similarly, to *lower an index* we can write $x_\mu = \eta_{\mu\nu} x^\nu$, where $\eta_{\mu\nu}$ is the same matrix as above ([36]).

The upper indices are not exponents but are indices of coordinates, coefficients or basis vectors. That is, in this context x^2 should be understood as the second component of x^μ rather than the square of x (this can occasionally lead to ambiguity). Note that the superscript index μ takes four values $\mu \in \{0, 1, 2, 3\}$ ([37]), one for each element

$$x^0 = ct, x^1 = -x, x^2 = -y, x^3 = -z.$$

Finally, using the index notation, the Einstein summation convention ([38]) (implicit summation on repeated indices), and the Minkowski metric $\eta^{\mu\nu}$, we can write the spacetime interval in a more compact form:

$$ds^2 = \Delta x_\mu \Delta x^\mu = \Delta x^\mu \Delta x_\mu := \sum_{\mu=0}^{3} \Delta x_\mu \Delta x^\mu.$$

Note that we can use any letter to indicate a summed index: $\Delta x_\mu \Delta x^\mu$ and $\Delta x_\alpha \Delta x^\alpha$ are exactly the same.

The Minkowski metric $\eta^{\mu\nu}$ is also called *Minkowski tensor*. In mathematics, a tensor is an algebraic object that describes a (multilinear) relationship between sets of algebraic objects related to a vector space ([39]). Objects that tensors may map between include vectors and scalars, and, recursively, even other tensors ([W11]).

---

[36] In physicists' terminology, a vector whose components are labelled with upper indices is called a *contravariant* vector whereas a vector whose components have lower indices is called *covariant*. Covariant vectors are in fact slightly different objects, they are the dual vectors (or sometimes called *covectors*), i.e. they are elements of the dual space. Inner products are only between contravariant and covariant vectors (or tensor) components. Since Minkowski space (as a vector space) is isomorphic to its dual space, we will simply treat covariant and contravariant vectors on an equal footing. Keep in mind however, that x_μ and x^μ are *different* objects. ([W3]).
A detailed discussion of this topic would lead us too far astray here.

[37] The symbol '∈' means 'belongs to' or 'is a member of'.

[38] Einstein summation convention ([E2], [W11]) is a convenient notation when manipulating expressions involving vectors, matrices, or tensors in general. The 'rules' of summation convention are:

- Each index can appear at most twice in any term.
- Indices which repeat in an expression are always summed over. This is known as *contraction*. Greek indices like μ, ν, α are always summed from 0 to 3.

For example, $A^{\mu\nu} v_\nu := \sum_{\nu=0}^{3} A^{\mu\nu} v_\nu$ is a valid expression – it is just left-multiplication of vector v_ν by 4×4 matrix $A^{\mu\nu} = [A^{\mu\nu}]$. Technically, one should write $A^\nu_\mu v_\nu$. However, using upper and lower indices on the same object makes expressions difficult to read, so it is common lower or raise all the indices. Contracting indices is just a notational convention, not a deep property of mathematics ([S1]).

[39] A *vector space* (also called, interchangeably, a *linear space*) is a collection of objects called vectors, which may be added together and multiplied ('scaled') by real or complex numbers, called *scalars*. The operations of vector addition and scalar multiplication must satisfy certain requirements. When the scalars are the real numbers, the vector space is called a *real vector space*, and when the scalars are the complex numbers, the vector space is called a *complex vector space*. A *complex number* is an element of a number system that extends the real numbers. Every complex number z can be expressed in the form $z = x + iy$, where x and y are real numbers (i.e. $x, y \in \mathbb{R}$) and i, called the *imaginary unit*, satisfies the equation $i^2 = -1$. The set of complex numbers is denoted by $\mathbb{C}$.
A set B of vectors in a vector space V is called a *basis* (pl.: *bases*) if every element v of V may be written in a unique way as a finite linear combination of elements of B: $v = \lambda_1 b_1 + ... + \lambda_n b_n$, where λ_i are scalars and $b_i \in$ B for $i = 1, 2, ...,$ n. A vector space that has a finite basis is called *finite-dimensional*. ([W11]).

The Minkowski tensor $\eta^{\mu\nu}$ is a *rank-2* tensor, which means that it assigns a scalar (i.e. number) to a pair of vectors – see e.g. (3.3). Exactly speaking, a rank-2 tensor is not just a matrix. The Minkowski tensor is only represented by the matrix $\eta^{\mu\nu}$, but that representation depends on the choice of coordinates.

Tensors can be used to describe physical properties, just like scalars and vectors. In fact tensors are merely a generalisation of scalars and vectors; a scalar is a rank-0 tensor, and a vector is a rank-1 tensor (see [D4] for more details).

3.2. Lorentz transformations

In 1905, the 'annus mirabilis' of physics, Einstein publishes four ground-breaking articles in *Annalen der Physik*. The first of those papers ([E3]) provides an interpretation of the photo effect with the hypothesis of light quanta. In the second paper ([E4]) Einstein develops a quantitative formulation of Brownian motion. Finally, in the last two papers ([E5, E6]) of his annus mirabilis Einstein formulates the special theory of relativity (SR) ([40]) that fundamentally changed our perceptions of space and time. ([E1])

All experimental observations show that our world is relativistic. Consequently, our theories of the fundamental interactions must be compatible with SR. In the physicist's terminology, those theories must be the same in all inertial frames. In order to ensure it, we require the equations in question to be covariant under Lorentz transformations – that is, they must have the same form in the two different frames ([41]).

An event, i.e. a point of spacetime, can be labelled by the values of its coordinates in the rest frame or by its coordinates in a moving frame. These are two different descriptions of a single event. The question is now, how do we go from one description to the other? In other words, what is the coordinate transformation relating the rest frame coordinates to the coordinates of the moving frame?

Since the structure of spacetime is non-trivial, one has to be careful when switching coordinate systems. Allowed transformations need to respect the laws of special relativity and they must leave the spacetime interval ds^2 unchanged. It means that they leave the scalar product in Minkowski space $a_\mu b^\mu := a_\mu \eta^{\mu\nu} b_\nu$ unchanged. It is what is called a *Lorentz scalar* ([S2]). It is important to understand that the scalar product of two vectors a_μ and b_μ is often written in many equivalent forms

$$a\cdot b = a_\mu b^\mu = a_\mu \eta^{\mu\nu} b_\nu = a_0 b^0 + a_i b^i = a_0 b_0 - \sum_{i=1}^{3} a_i b_i = a_t b_t - (a_x b_x + a_y b_y + a_z b_z) = a_t b_t - \boldsymbol{a}\cdot\boldsymbol{b}. \tag{3.4}$$

It is obvious that the (squared) four-dimensional length we described above can be written as $a^2 = a_\mu^2 := a_\mu a^\mu = a_t a_t - \boldsymbol{a}\cdot\boldsymbol{a} = a_t^2 - \mathbf{a}^2$.

One important consequence of the non-trivial spacetime structure is that observers who move relative to each other measure different time intervals between two events. In physical terms, this means that time appears delayed for the moving observer ([S2]).

It turns out that there are three kinds of allowed transformations: rotations, boosts, and translations.

- A *rotation* is a switch to a new coordinate system that is oriented differently with respect to the original coordinate system.

---

[40] The designation 'special relativity' is due to the fact that the distinction between space and time is not absolute, but 'relative'. The theory is considered 'special' as it only deals with reference frames moving at constant velocities.

[41] An equation is covariant if both sides transform the same way. This implies that the equation remains true after a Lorentz transformation. Sometimes, covariant equations are called invariant – see Section 3.6 for details. Lorentz transformations were actually known before Einstein. Henri Poincaré, who also suggested the name, gave Lorentz transformations their final form in 1905 shortly before the publication of SR. ([E1])

- A *boost* is a switch to a coordinate system that is moving with a different constant velocity with respect to the original coordinate system.
- A *translation* is a switch to a shifted coordinate system. Since we are dealing with spacetime, we can consider temporal shifts $t \to t + a$ or spatial shifts $\boldsymbol{x} \to \boldsymbol{x} + \boldsymbol{b}$.

Rotations only affect the spatial components of a 4-vector and can be described by three basis matrices defining rotations around the x-axis, y-axis, and z-axis, respectively. Any other rotation can be thought of as a combination of rotations around the three coordinate axes.

For example, the following matrix describes a rotation around the x-axis

$$R^{\mu\nu}(\theta) = R^{\mu}_{\nu}(\theta) := \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos(\theta) & -\sin(\theta) \\ 0 & 0 & \sin(\theta) & \cos(\theta) \end{bmatrix},$$

where θ is the angle of rotation ([42]).

Rotations, boosts, and all transformations that are possible by combining them are called *Lorentz transformations*. As mentioned above, the temporal component ct is never affected by rotations, only by boosts and temporal shifts (i.e. shifts to a different point in time). Thus, one can really say that the Lorentz boosts are the 'interesting' Lorentz transformations. The remainder is just rotations of our familiar 3-dimensional Euclidean space.

Moreover, we call rotations, boosts, translations, and all their combinations *Poincaré transformations*.

3.3. Four-vectors revisited

We called a vector $x_\mu = [ct \quad x \quad y \quad z]^T$ (here written as a transposed row vector) that describes the position in spacetime '4-vector'.

One confusing thing is that mathematicians use the word vector quite differently than physicists. For mathematicians, any element of a vector space is a vector. Thus for them, a 4-vector is just an element of a four-dimensional vector space. In contrast, physicists define mathematical objects like scalars, vectors, or tensors in terms of how they behave under coordinate transformations ([S2]).

Any set of four quantities that transforms from one Lorentz frame to another just like the coordinates ct, x, y, z of a point in spacetime is called a *4-vector* (or *four-vector*). The prototypical 4-vector is hence x_μ.

What does this actually mean? It can be best explained with an example.

Example 3.2. Let us consider the important *energy-momentum vector* $p_\mu := (E/c, \boldsymbol{p})$ of a particle with mass m_0 and velocity $\boldsymbol{u}$ in a Lorentz frame L with coordinates ct, x, y, z. Its spatial components are the three components of relativistic momentum 3-vector $\boldsymbol{p} = (p_x, p_y, p_z)$ whereas the first component p_t is the relativistic energy E divided by c, where $E = (\boldsymbol{p}^2c^2 + m_0^2c^4)^{1/2}$. Recall that $\boldsymbol{p}^2$ is the squared length of the 3-vector $\boldsymbol{p}$: $\boldsymbol{p}^2 := (p_x)^2 + (p_y)^2 + (p_z)^2$ ([43]). Let us show that p_μ is a 4-vector.

Assume that L' is a Lorentz frame with coordinate axes ct', x', y', z' moving uniformly with velocity $\boldsymbol{v} = (v, 0, 0)$ parallel to the x-axis of L. This boost transformation is represented by the matrix

[42] In Einstein notation, the usual element reference for the m^{th} row and n^{th} column of a matrix A becomes A^m_n.

[43] For the components of a 4-vector a_μ we will use the notation a_0, a_1, a_2, a_3 or a_t, a_x, a_y, a_z.

$$B^{\mu\nu} = B^{\mu\nu}(v) := \begin{bmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

where $\beta := v/c$ and $\gamma := 1/\sqrt{1-\beta^2}$. Then ([44]) $x'_\mu = B^{\mu\nu}x_\nu$ and it implies that

(3.5) $$ct' = \gamma(ct - \beta x), \quad x' = \gamma(x - \beta ct), \quad y' = y, \quad z' = z.$$

The vector p_μ in order to be a 4-vector has to transform like the position vector x_μ, i.e. in the frame L' the vector p_μ must be represented by

(3.6) $$E'/c = \gamma(E/c - \beta p_x), \quad p'_x = \gamma(p_x - \beta E/c), \quad p'_y = p_y, \quad p'_z = p_z.$$

Let us show, for example, that $p'_x = \gamma(p_x - \beta E/c)$.

Our particle is moving with velocity $\boldsymbol{u} = (u_x, u_y, u_z)$ in the frame L. The second frame L' is moving at velocity $\boldsymbol{v}$ with respect to the first frame, along the x-axis of L, and the particle has a velocity $\boldsymbol{u}' = (u'_x, u'_y, u'_z)$ as seen in the second frame. Recall that the (relativistic) momentum 3-vector $\boldsymbol{p}'$ is given (in the frame L') by

$$\boldsymbol{p}' = \frac{m_0 \boldsymbol{u}'}{\sqrt{1 - (u'/c)^2}},$$

where u' is the magnitude of the velocity $\boldsymbol{u}'$. Thus

$$p'_x = \frac{m_0 u'_x}{\sqrt{1 - (u'/c)^2}}.$$

Since $v_x = v$, the velocity addition law gives us:

$$u'_x = \frac{u_x - v}{1 - \frac{u_x v}{c^2}} = \frac{u_x - \beta c}{1 - \frac{u_x}{c}\beta}.$$

Of course, $u'_y = u_y$ and $u'_z = u_z$. Using these equalities it is easy to check that

$$\frac{1}{\sqrt{1 - (u'/c)^2}} = \gamma \frac{1 - \left(\frac{u_x}{c}\right)\beta}{\sqrt{1 - (u/c)^2}}$$

Consequently

$$p'_x = \frac{m_0 u'_x}{\sqrt{1 - (u'/c)^2}} = m_0\gamma \frac{u_x - \beta c}{1 - (u_x/c)\beta} \frac{1 - (u_x/c)\beta}{\sqrt{1 - (u/c)^2}} = \gamma \frac{m_0 u_x - \beta m_0 c}{\sqrt{1 - (u/c)^2}} = \gamma(p_x - \beta E/c),$$

because the relativistic energy E is given by

$$E = \frac{m_0 c^2}{\sqrt{1 - (u/c)^2}}.$$

Comparing (3.6) with (3.5) we infer that $p'_\mu = B^{\mu\nu}(v)p_\mu$. Thus the components of the energy-momentum vector p_μ transform exactly like the coordinates ct, x, y, z. The same holds for all other Lorentz transformations and it implies that $p_\mu = (E/c, \boldsymbol{p})$ is a 4-vector ([45]).

---

[44] Recall that $x'_\mu = B^{\mu\nu}x_\nu := \sum_{\nu=0}^{3} B^{\mu\nu}x_\nu$. This is 'four equations in one', since $\mu = 0,1,2,3$.

[45] Mind that a 4-vector is *not* just some one-dimensional array of four numbers. And it is *not* just a matter of adding any fourth t-component to a 3-vector. Feynman [F5] gives the following example: consider the velocity 3-vector $\boldsymbol{v}$ with components $v_x = dx/dt$, $v_y = dy/dt$ and $v_z = dz/dt$. Then $v_\mu = (d(ct)/dt, dx/dt, dy/dt, dz/dt) = (c, \boldsymbol{v})$ is *not* a 4-vector. The correct 4-velocity is $v_\mu = \gamma(c, \boldsymbol{v})$, where $\gamma = dt/d\tau$ and τ is the proper time.
Notice, however, that this vector is not a velocity vector in the conventional sense. It determines rather a kind of a direction in the four-dimensional geometry. ([T5])

Consequently, Lorentz transformations leave the expression $p_\mu p^\mu = (E/c)^2 - \boldsymbol{p}^2 = m_0^2 c^2$ invariant, which means that the mass m_0 is invariant under Lorentz transformations. The formula $m_0^2 c^2 = (E/c)^2 - \boldsymbol{p}^2$ is famously known as the *relativistic energy-momentum relation*.

Since the scalar product $a_\mu a^\mu$ ([46]) is invariant under Lorentz transformations for a general 4-vector a_μ, it gives a powerful method of calculating kinematical variables and their transformations from one inertial frame to another ([47]).

In physical terms, this means that Lorentz transformations describe changes between frames of reference that respect the postulates of special relativity (SR).

Since the Lorentz transformations are linear ([48]), the *sum* and *difference* of 4-vectors is also a 4-vector. New 4-vectors may also be obtained by *multiplying/dividing* by a scalar invariant, such as the proper time interval $d\tau$ or the mass m_0.

3.4. Fields

The most important types of fields for our topics are scalar fields, vector fields, and spinor fields ([S2]) ([49]).

A *scalar field* is a mapping $x_\mu \to s(x_\mu)$ that assigns to each spacetime point x_μ a (real or complex ([50])) number $s(x_\mu)$. As an example of a scalar field, consider a solid block of material that has been heated at some places and cooled at others, so that the temperature of the body varies from point to point in a complicated way. Of course, the temperature may change in time. Then the temperature will be a function of time and x, y, and z, the position in space measured in a rectangular coordinate system. Temperature is a scalar field ([F5]). Another example is the Higgs field (see Section 7.3).

Similarly, a *vector field* is a mapping $x_\mu \to v(x_\mu)$, where v is a vector (i.e. element of a vector space). We shall mostly consider 3- and 4-vector fields, i.e. mappings $x_\mu \to \boldsymbol{v}(x_\mu)$ and $x_\mu \to v_\mu(x_\mu)$ ([51]), respectively. As an example, consider a rotating body. The velocity of the material of the body at any point is a 3-vector which is a function of position (and time). Other examples include 4-vector gauge fields, describing spin-1 particles such as the photon (see Section 5.4).

When talking about quantum matter fields we will need the third kind of field which is known as *spinor field*. A *spinor field* is a mapping $x_\mu \to \psi(x_\mu)$ assigning to each spacetime point x_μ a spinor $\psi(x_\mu)$. Since there are different kinds of spinors, there are also different kinds of spinor fields. Every fermionic field (i.e. a quantum field whose quanta are fermions: electrons, quarks, etc.) is a spinor field.

Spinors are (unintuitive) two-component objects that live somewhere between a scalar and a 4-vector as far as their behaviour under coordinate transformations is concerned. When we

---

[46] Recall that the (squared) length a^2 of a 4-vector a_μ is given by the scalar (i.e. inner Minkowski) product of the vector a_μ with the covector a^μ: $a^2 = a_\mu a^\mu$.

[47] The Lorentz transformations are sometimes defined as those transformations that leave the scalar product of Minkowski spacetime invariant – see Section 3.5.

[48] A *linear transformation* is a mapping $T: V \to W$ between two vector spaces that preserves the operations of addition and scalar multiplication: $T(\alpha_1 v_1 + \alpha_2 v_2) = \alpha_1 T(v_1) + \alpha_2 T(v_2)$, where $v_1, v_2 \in V$ and α_1, α_2 are scalars.

[49] These fields are important in particle physics. General relativity requires tensor fields ([F8]).

[50] Recall that every complex number z can be expressed in the form $z = x + iy$, where x and y are real numbers and i, called the *imaginary unit*, satisfies the equation $i^2 = -1$. To avoid confusion, in this paper the imaginary unit i is written in roman typeface, while an index *i* is written in italics.

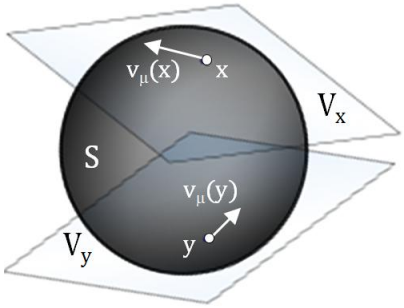
[51] Here again we use this standard but somewhat confusing notation for 4-vectors: the same symbol is used for the whole vector and its components.

want to calculate how a spinor looks like after a rotation or boost, we can no longer use the 4×4 matrices as for 4-vectors. Instead, we need 2×2 matrices.

We shall discuss spinors later in Section 5.8. But from a mathematical viewpoint spinors fields are not special. For our purposes, a field F is just a mapping that assigns to each spacetime point x_μ a mathematical object $F(x_\mu)$. For example, we will consider below a matrix field $x_\mu \rightarrow M(x_\mu)$, when talking about Yang-Mills theory.

There is still one aspect of fields that we need to talk about before we can discuss later how fields evolve in spacetime. The field values live in an abstract space 'on top' of spacetime. A field is an object that glues this abstract space and spacetime together. It is important to keep in mind that there is a copy of the basic field space above each spacetime point. This total field space is all these individual basic field spaces taken together ([S2]) ([52]).

For example, the values v_μ of a vector field $x \rightarrow v_\mu(x)$ do not belong to the **same** vector space V. At each point x of spacetime, a different copy V_x of a basis ('generic') vector space V is attached (i.e. each V_x is isomorphic to V ([53])). This means that the mapping $x \rightarrow v_\mu(x)$ assigns a vector $v_\mu(x)$ in the space V_x to each point x of spacetime. Consequently, one cannot, e.g. just simply add vectors $v_\mu(x) + v_\mu(y)$ (for $x \neq y$), because they belong to **different** vector spaces ([54]). The easiest example to think of is a sphere S in $\mathbb{R}^3$ with the tangent planes V_x attached to each point $x \in S$:



Another important observation is that there is a different field space for each field. For example, temperatures and the Higgs field occupy different field spaces even though both are mathematical scalar fields.

3.5. Symmetries and groups

Although the concept of symmetry is familiar to most people, its mathematical definition is not that obvious and the symmetries we shall encounter in this article are rather abstract. The understanding of symmetries may be one of the biggest obstacles in comprehension of Yang-Mills theory ([B3]).

When physicists talk about symmetry, they mean something particular. Symmetry does mean a different thing for physicists than for members of the public. Physicists understand 'symmetry' to be transformations that leave the object under study unchanged or 'invariant' as they usually express it. In physics the object to remain invariant is usually the laws of physics, i.e. the considered equations (typically the Lagrangian density – see Section 5.1 for details). Though

---

[52] In technical terms, the total field space is a fibre bundle with spacetime as its base space and basic field spaces as fibres.

[53] Two vector spaces V and W are said to be *isomorphic* if there exists an one-to-one linear transformation (called *isomorphism*) T from V onto W.

[54] Even though all vector spaces V_x are isomorphic to V, there is no canonical isomorphism between them. This is an issue to deal with when considering e.g. the derivative of a vector field. To describe the differential evolution equation of a field one has to define a connection in the fibre bundle.

this definition is technically equivalent with our every day intuition of what symmetry is, it is formulated in a rather more abstract fashion and thus takes some time getting used to. ([B3])

The concept of symmetry, i.e. the invariance of a theory under certain transformations of the quantities contained in it, plays an important, if not the most important role, in the mathematical formulation of the laws of nature.

The symmetries of nature determine the things that remain constant, i.e. are conserved. Those are the guideposts in physics, the quantities like energy and momentum. For instance, energy is conserved, we now understand, because there is a symmetry of nature that tells us the laws of physics do not change over time ([M10]).

For every global continuous symmetry – i.e. a transformation of a physical system that acts the same way everywhere and at all times – there exists an associated time independent quantity: a conserved 'charge'. This connection went unnoticed until 1918 when Emmy Noether proved her famous theorem relating symmetry and conservation laws ([55]). Thus due to the invariance of the laws of physics under spatial transformations, momentum is conserved. Due to time translational invariance, energy is conserved. And due to the invariance under a change in phase of the wave functions of charged particles, electric charge is conserved ([T2]).

When considering the role of symmetry in physics from a historical viewpoint, it is worth keeping in mind the distinction between implicit and explicit uses of the notion. Symmetry considerations have always been applied to the description of nature, but for a long time in an implicit way only. The real turning point in the use of symmetry in science came, however, with the introduction of the group concept and with the ensuing developments in the theory of transformation groups ([B7]).

This is because the group-theoretic definition of symmetry as *invariance under a specified group of transformations* allowed the concept to be applied much more widely, not only to spatial figures but also to abstract objects such as mathematical expressions – in particular, expressions of physical relevance such as dynamical equations ([F6]).

A *group* is defined to be a set G with a composition rule $\bullet$: $G \times G \to G$ that combines any two elements g and h of G to form an element of G, denoted $g \bullet h$ (or simply gh), such that the following three requirements are satisfied ([A1], [B2], [H0], [W11]):

(G1) *Associativity*

$$(gh)f = g(hf) \quad \text{for all } g, h, f \in G.$$

(G2) *Identity element*

There is an *identity* element $1 \in G$ such that

$$1g = g1 = g \quad \text{for all } g \in G.$$

(G3) *Inverse element*

For each element $g \in G$ there exists an element $h \in G$ such that

$$gh = hg = 1.$$

Such an element is unique, it is called the *inverse* of g and denoted g^{-1}.

The map $G \times G \to G$ is called the *product operation* for the group. Part of the definition of a group G is that the product operation map $G \times G$ into G, i.e., that the product of two elements of G be again an element of G.

Many physical and mathematical objects or physical theories possess symmetry. If the set of given symmetries forms a group, then we can compose them. We usually interpret gh as 'act with h first and then act with g'. (G1) means that ghf does not need brackets because it implies:

---

[55] Informally stated, Noether's theorem says that to every continuous symmetry of a theory there corresponds a conservation law and vice versa. Amalie Emmy Noether (1882 – 1935) was a German mathematician who made many important contributions to abstract algebra.

act with f then h then g. (G2) says that doing nothing is a symmetry, the identity 1. This guarantees that the set G is not empty. (G3) means that a symmetry transformation g can be reversed, which gives the inverse g^{-1}. The inverse is itself a symmetry. The conclusion is that group theory is the mathematical framework of symmetry ([M2]).

Usually one considers the concept of group as a generalization of the multiplication of numbers. However, in a general group the law of combination need not be commutative, i.e. $gh \neq hg$. If it is commutative ($gh = hg$), the group is *Abelian;* if not, it is *non-Abelian* ([56]).

Example 3.3. Let O(1,3) denote the set of all real-valued 4×4 matrices A with the property

$$A^T \eta^{\mu\nu} A = \eta^{\mu\nu},$$

where $\eta^{\mu\nu}$ is the Minkowski metric (see Section 3.1) and A^T denotes the transpose ([57]) of the matrix A. Matrices with this property are called *orthogonal*. This is the reason for 'O' in O(1,3). The numbers 1 and 3 refer to the signature of the Minkowski metric. We will show that O(1,3), with the ordinary matrix multiplication as the 'product', is a group. Since matrix multiplication is associative, the condition (G1) is satisfied.

Let us take any two matrices A, B ∈ O(1,3). Then $A^T \eta^{\mu\nu} A = \eta^{\mu\nu}$ and $B^T \eta^{\mu\nu} B = \eta^{\mu\nu}$. Since $(AB)^T = B^T A^T$ we infer that $(AB)^T \eta^{\mu\nu}(AB) = (B^T A^T)\,\eta^{\mu\nu}(AB) = B^T(A^T \eta^{\mu\nu} A)B = B^T \eta^{\mu\nu} B = \eta^{\mu\nu}$. Consequently, AB ∈ O(1,3).

Let I_4 denote the 4×4 *identity matrix*

$$I_4 := \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Of course, $I_4 A = A I_4 = A$ for any 4×4 matrix A. Moreover ${I_4}^T \eta^{\mu\nu} I_4 = I_4 \eta^{\mu\nu} I_4 = \eta^{\mu\nu}$. Thus the identity I_4 belongs to O(1,3).

Now let us take any element A ∈ O(1,3). Then $A^T \eta^{\mu\nu} A = \eta^{\mu\nu}$ and multiplying this equality by $\eta^{\mu\nu}$ we get $\eta^{\mu\nu} A^T \eta^{\mu\nu} A = \eta^{\mu\nu}\eta^{\mu\nu} = I_4$. Putting $A^{-1} := \eta^{\mu\nu} A^T \eta^{\mu\nu}$ we obtain $A^{-1} \in O(1,3)$ and $A^{-1}A = AA^{-1} = I_4$. Consequently, A^{-1} is the inverse of A, i.e. the condition (G3) is satisfied.

The group O(1,3) is not Abelian. Indeed, let ([58])

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}.$$

It is easy to show that A, B ∈ O(1,3). Let us compute AB and BA

$$AB = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

$$BA = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & -1 & 0 & 0 \end{bmatrix}.$$

Thus AB ≠ BA and it means that O(1,3) is not Abelian.

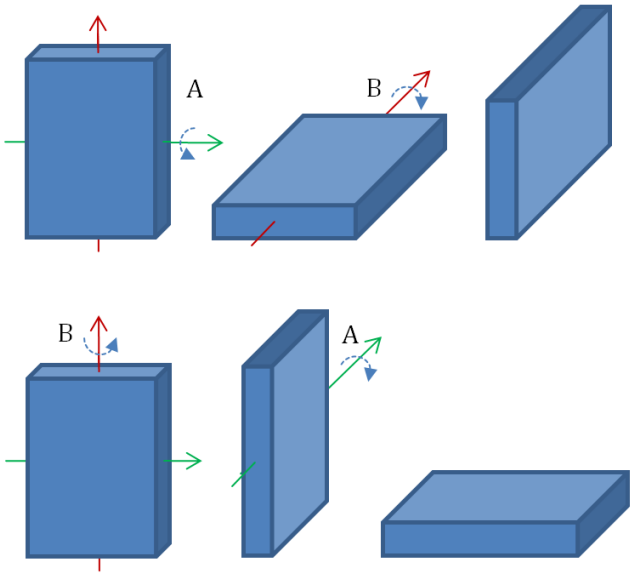
---

[56] Abelian groups are named after Norwegian mathematician Niels Henrik Abel (1802 – 1829).

[57] The transpose A^T of a matrix $A = [a_{ij}]$ is a new matrix whose rows are the columns of the original, i.e. $A^T = [a_{ji}]$.

[58] A and B are rotation matrices around the x and y axis by $\pi/2$, respectively – see Section 3.2.

Notice that A and B are rotation matrices around the x and y axis by $\pi/2$, respectively (see Section 3.2). It is evident from the following illustration that rotations do not commute:



In technical terms, Lorentz transformations are defined as linear transformations $L: \mathbb{M} \rightarrow \mathbb{M}$ ([59]) that preserve the inner product of Minkowski space $\mathbb{M}$. Equivalently, every Lorentz transformation can be represented as a real-valued 4×4 matrix Λ with the property $\Lambda^T \eta^{\mu\nu} \Lambda = \eta^{\mu\nu}$, i.e, $\Lambda \in O(1,3)$. Thus we can identify the set of all Lorenz transformations with the group $O(1,3)$, which is therefore called *Lorentz group*.

The *Poincaré group* consists of the Lorentz transformations and translations T in four dimensions. Consequently, the Lorentz group is a subgroup of the Poincaré group ([60]). The Poincaré group is the group of special relativity because the assumption of homogeneity of spacetime requires the invariance of laws of physics under 4-dimensional translations.

The application of the theory of groups and their representations for the exploitation of symmetries in the quantum mechanics of the 1920s undoubtedly represents the second turning point in the twentieth-century history of physical symmetries. It is, in fact, in the quantum context that symmetry principles are at their most effective. Wigner and Weyl ([61]) were among the first to recognize great relevance of symmetry groups to quantum physics and the first to reflect on the meaning of this ([B7]).

When we say that nature is invariant under some symmetry, it means

- all objects in the theory have well-defined transformation properties (i.e. well defined 'representation' of the symmetry group) under the symmetry, and
- every interaction is invariant under the symmetry transformations.

---

[59] Reminder: $L: \mathbb{M} \rightarrow \mathbb{M}$ is linear if for any two vectors $x, y \in \mathbb{M}$ and any scalar s the following two conditions are satisfied: $L(x + y) = L(x) + L(y)$ and $L(sx) = sL(x)$.

[60] The Poincaré group was later called 'inhomogeneous Lorentz group' by Eugene Wigner in his fundamental paper [W10]. The word 'inhomogeneous' was added because the group takes into account spacetime translations in addition to Lorentz transformations. In the mathematical language, this inhomogeneous group is a 'semi-direct product' of the spacetime translation group and the four-dimensional Lorentz group ([K1]). The Poincaré group is the isometry group of the Minkowski space $\mathbb{M}$.
Eugene Paul "E. P." Wigner (1902 – 1995) was a Hungarian-American theoretical physicist. He received the Nobel Prize in Physics in 1963.

[61] Hermann Klaus Hugo Weyl (1885 – 1955) was a German mathematician, theoretical physicist and philosopher.

The 'objects' of particle physics are particle fields. The interactions in particle physics are the operators in the Lagrangian (see Section 5.1). From a mathematical viewpoint, *particle* is an object that has a well-defined transformation property under Lorentz symmetry ([W2]).

An important distinction in the study of symmetries in physics is the one between *external* and *internal* symmetries (see e.g. [F15]). *External* symmetries are coupled to the Poincaré group, i.e. they are the symmetries of spacetime.

However, in field theory, new symmetries also appear which do not have an analogue in the classical mechanics of particles. We shall see later that the symmetries of a theory are not limited to the invariance under coordinate transformations. At least as important is the class of *internal* symmetries (so-called *gauge symmetries*), i.e. transformations unrelated to the spacetime coordinates of the system.

These are the internal symmetries, which will be discussed below in the context of Yang-Mills theory. Internal symmetries act on fields, not directly on spacetime. That is, they work in mathematical spaces that are generated by the fields.

As we shall see, internal symmetries are symmetries that arise in the Lagrangian because fields appear in a symmetric way, e. g. a complex scalar field ψ can be invariant under the global phase shift $\psi \rightarrow e^{i\alpha}\psi$. In the group theoretical language, this symmetry is described by U(1) ([62]). These symmetries are internal in the sense that they do not 'see' the Poincaré group ([63]).

Another important distinction in physics is between *global* and *local*. *Global symmetries* are transformations that leave the physics unchanged and apply in the same way in all of spacetime. Examples are Lorentz transformations.

A system that has *local symmetries* is invariant under transformations that change from point to point, i.e. different transformations are carried out at different individual spacetime points. An example of such a transformation is a deformation in which every point is translated, but by a different amount.

The global symmetries are found to be associated with properties of particles, e. g., whether they are matter or antimatter, whether they carry electric charge, and so on. Local symmetries are found to be associated with forces. In fact, all the fundamental interactions of nature are related to very special local symmetries. The latter class of symmetries goes under the name of *gauge symmetries*. The work of C.N. Yang and R. Mills reveals the significance of these symmetries. ([M0])

To illustrate how a gauge symmetry leads to interactions, we shall look at a toy model describing a simplified financial market (see [B3], [S3a] and references therein).

Example 3.4. Let us assume that our simplified financial market consists of several countries and the basic process we try to describe is that money and goods can be traded and carried around. For concreteness, we call the countries Germany, Czech Republic and Poland, and the currencies Euro (EUR) in Germany, Czech Crown (CZK) in Czech Republic, and Zloty (PLN) in Poland.

In an isolated country it is in principle possible to rescale all monetary values with any factor and nothing in the economy would change. If we for example multiply all prices with 10 everything will look very expensive, but if we at the same time multiply all salaries, savings and loans with the same number nothing really changes (we pretend this rescaling would work perfectly). A crucial observation is now that we deal with a global symmetry since the absolute value of fiat money is, in general, not determined. We can shift the currency or alternatively, all prices without any physical effect.

---

[62] U(1) is the simplest internal symmetry group – see Section 5.5 for details.

[63] In the mathematical language, this means that the generators of an internal group commute with all generators of the Poincaré group.

In the real world, different countries scale their economy differently. How then to deal with Euro when we come e.g. to Czech Republic? This is only possible if we introduce bookkeepers which keep track of the values of the local currencies and are able to exchange one currency for another. We can then imagine that the bookkeepers always adjust their exchange rates perfectly whenever the value of a local currency changes. With the exchange rates in place we see that our scaling symmetry is now local, we can scale the economy of every country independently of the other countries. We can therefore say that as soon as we introduce bookkeepers, our system is invariant under local transformations. It is conventional to call this invariance gauge symmetry.

So far, our bookkeepers are purely mathematical ingredients which we introduced to make our description invariant under local transformations. In our finance example, it is easy to imagine that bookkeepers can influence the dynamics of a system and even become dynamical actors on their own. We can imagine that there are imperfections in the exchange rates, i.e. that the exchange rates fluctuate. If this is the case, a trader can buy and sell currencies since this can be a lucrative business. And this is an explicit example of how our bookkeepers can influence the dynamics of the system.

For example, let us imagine the exchange rates are as follows EUR/CZK = 30, CZK/PLN = 0.2 and PLN/EUR = 0.3. Now a trader is able to earn money simply by trading currencies. If he starts with 1 EUR, he can trade it for 30 CZK, then use it to buy 6 PLN, and finally trade these for 1.8 EUR. This means that the trader has more money than he started with and if this is the case, investors will travel this circle over and over again to make money. Consequently, fluctuations in the exchange rates produce a 'force' that makes investors move around in this circle.

In this example we can also see that the interaction goes both ways, we know from the real world that movements of money between countries will change the exchange rates. In other words, the exchange rates affect the money and the money affects the exchange rates.

Promoting bookkeepers (exchange rates) to dynamical objects which follow their own rules turns our model of the financial market into a gauge theory with the exchange rates as a 'gauge field'. It means that exchange rates are adjusted dynamically depending on what else happens in the system.

In general, as we shall see, a theory that is globally invariant will not be invariant under locally varying transformations. The invariance under gauge transformations requires the introduction of gauge vector fields, which are interpreted as the quanta mediating the interactions among the fermions that are the fundamental constituents of matter. Gauge symmetries leave certain quantum numbers such as spin or quark colour unchanged. As an example, we will discuss in Chapter 7 the isospin symmetry group SU(2) and the colour symmetry group SU(3), which are of great importance for the physics of weak and strong interactions.

We shall see all these things more clearly when we go into more detail, but the important conceptual point to be grasped is this: one may view these special force fields and their interactions as existing in order to permit certain local invariances to be true. ([A1])

3.6. Invariance and covariance

Before we move on, we have to clarify two important notions, which we have already come upon above. Firstly, we call something *invariant*, if it does not change under transformations.

For instance, let us consider a quantity like Q = Q(A, B, C, ...) that depends on other quantities A, B, C, ... If we transform A, B, C, ... → A', B', C', ... and we have Q(A', B', C', ...) = Q(A, B, C, ...) then Q is called *invariant* under this transformation. We can express this differently using the word symmetry. *Symmetry* is defined as invariance under a transformation or class of transformations.

For example, the spacetime interval between two events remains the same, i.e. invariant, under Lorentz transformations. In the same way, the scalar product $a_\mu b^\mu$ of any two 4-vectors is also invariant under Lorentz transformations.

Covariance means something similar, but may not be confused with invariance. An equation is called *covariant*, if it takes the same **form** when the objects in it are transformed. For example, as we shall see in Section 4.3, the Maxwell equations of electrodynamics are *Lorentz covariant*, the word 'covariant' here meaning precisely that both sides of an equation transform in the same way (i.e. consistently) under Lorentz transformations ([S2]).

In other words, saying that objects are Lorenz invariant means that they do not depend on our Lorentz frame at all, while objects being Lorentz covariant means that they do change in different frames, but precisely as the Lorentz transformation dictates ([S1]).

Confusingly enough, this use of the word 'covariant' is evidently quite different from the one encountered previously in an expression such as 'a covariant 4-vector', where it just meant a 4-vector with a downstairs index (see footnote ([36])). This new meaning of 'covariant' is actually more accurately captured by an alternative name for the same thing, which is 'form invariant'.

Why is this idea so important? Consider the special relativity principle, which states that the laws of physics should be the same in all inertial frames. The way in which this physical requirement is implemented mathematically is precisely via the notion of *covariance under Lorentz transformations*. For, consider how a law will typically be expressed. Relative to one inertial frame, we set up a coordinate system and describe the phenomena in question in terms of suitable coordinates, and such other quantities (forces, fields, etc) as may be necessary. We write the relevant law mathematically as equations relating these quantities, all referred to our chosen frame and coordinate system. What the relativity principle requires is that these relationships – these equations – must *have the same form* when the quantities in them are referred to a different inertial frame ([A1]).

Note that we must say 'have the same form', rather than 'be identical to', since we know that coordinates, at least, are not identical in two different inertial frames. This is why the term 'form invariant' is a more helpful one than 'covariant' in this context, but the latter is more commonly used ([A1], [S2]).

3.7. Natural units

In particle physics, a widely adopted convention is to work in a system of units, called *natural units*, in which the speed of light is set equal to unity ([A1], [M1]):

$$c = \hbar = 1,$$

where $\hbar$ (called h–bar) ([64]) is the reduced Planck's ([65]) constant introduced by Dirac ([66]) for $h/2\pi$, where h is the Planck's constant. This avoids having to keep track of untidy factors of $\hbar$ and c throughout a calculation; only at the end is it necessary to convert back to more usual units.

To understand the meaning of these units, observe first of all that $\hbar$ and c are universal constants, i.e. they have the same numerical value for all observers. The speed of light has the value $c = 299{,}792{,}458$ m/s, but instead of using the meter, we can decide to use a new unit of length (or a new unit of time) defined by the statement that in these units $c = 1$.

The Planck's constant (or Planck constant) h is another universal constant . It is the quantum of action ([67]) – the fundamental quantity in quantum mechanics that sets the scale for quantum

[64] The numerical value of $\hbar$ is 1.055×10^{-34} m^2kg/s.

[65] Max Karl Ernst Ludwig Planck (1858 – 1947) was a German theoretical physicist – 1918 Nobel Prize in Physics.

[66] Paul Adrien Maurice Dirac (1902 –1984) was an English theoretical physicist who is regarded as one of the most significant physicists of the 20th century – 1933 Nobel Prize in Physics.

[67] Roughly speaking *action* is a physical quantity which is equal to the kinetic energy, minus the potential energy, integrated over time [F4].

mechanical effects. It gives the ratio of the energy of a photon to its frequency, and by the mass-energy equivalence, the relationship between mass and frequency. By choosing units such that c $= 1$, units of mass and length retain their SI definitions ([68]) in terms of kilograms and meters, but time is transformed into a length and velocity is dimensionless.

In particle physics a useful unit of energy is the *electron-volt* (eV) equal to the energy gained by an electron when the electrical potential at the electron increases by one volt (notice that eV is not an SI unit). The electron-volt equals 1.602×10^{-19} joule. The abbreviation MeV indicates 10^6 (1,000,000) electron-volts; GeV, 10^9 (1,000,000,000); and TeV, 10^{12} (1,000,000,000,000). ([B9]). By mass-energy equivalence $E = mc^2$, the electron-volt corresponds to a unit of mass. It is common in particle physics, where units of mass and energy are often interchanged, to express mass in units of eV/c^2, or in terms of eV using natural units with c set to 1. An electron, for instance, has a mass of 0.511 MeV, that is about 10^{-30} kg. A proton has a mass of 938 MeV, which is 1,836 times the mass of an electron.

A typical length-scale in particle physics is the *fermi:* 1 fm $= 10^{-15}$ m – the radius of a proton. Then, in natural units, 1 fm $\approx 1/200$ MeV.

4. Classical electrodynamics as a gauge theory

In a famous 1954 paper [Y1], C.N. Yang and R.L. Mills proposed a broad class of classical field theories, which are known today as *Yang-Mills theories*. Inspired by Maxwell's theory of electromagnetism, Yang and Mills studied a more general (non-Abelian) gauge symmetry.

When quantised, those Yang-Mills theories became the mainstay for developments in particle physics in the second half of the twentieth century. As noted above, examples of quantised Yang-Mills theories include many of our most important and successful physical theories, including quantum electrodynamics (QED), the electroweak theory, the standard model of particle physics (SM), and the GUTs – grand unified theories. ([N2])

Although general relativity (GR) also satisfies a 'gauge symmetry' (principle of general covariance), it is not known whether it is possible to cast it as a Yang-Mills theory. This is rather unfortunate because quantisation of Yang-Mills theories is well understood, but not of general gauge theories ([69]) ([N2]).

However, what is today understood by the term 'Yang-Mills Theory' is very far from the original formulation of Yang and Mills. In 1954, Yang and Mills looked at a very special problem related to so-called 'isospins' and did not formulate the theory in the abstract language of 'principal fibre bundles', which is the standard way today. ([70])

We begin with a brief review of classical electrodynamics, which is a simple example of a gauge theory (see [A1], [F2–F6] and [S17] as a general reference for this chapter). The particular local invariance relevant to electromagnetism is the *gauge invariance* of Maxwell equations: in the quantum form of the theory (i.e. in quantum electrodynamics QED), this property is directly

---

[68] The International System of Units (SI) is based on the meter-kilogram-second (MKS) system of units.

[69] From a mathematical viewpoint, the structure group (gauge group) in a Yang-Mills theory is a compact Lie group (see Section 7.2), which does not always have to be the case in a general 'gauge theory'. For example the structure group Diff($\mathbb{M}$) (= group of diffeomorphisms on the spacetime manifold $\mathbb{M}$) in general relativity is not even locally compact. However, a relationship of Yang-Mills theory with the gauge description of general relativity remains an active area of research.

[70] Gauge fields are defined on principal fibre bundles as connection 1-forms with values in the Lie algebra of the gauge group. These fields correspond in the related quantum field theory to gauge bosons. Matter fields (i.e. fermionic fields) are introduced using 'associated vector bundles'. Connections (the gauge fields) define a covariant derivative on these associated vector bundles, leading to a coupling between gauge fields and matter fields ([H2]).

This highly abstract formalism is presumably the reason that for a considerable period of time the present author was unable to comprehend the meaning of Yang-Mills or, more generally, of a gauge theory in physical terms. It was only when he had gained at least a partial understanding of the concept of gauge symmetry in the theory of electromagnetism that he was finally able to grasp what this whole concept was all about.

related to an invariance under *local phase transformations* of the quantum fields. A generalized form of this phase invariance also underlies the theories of the weak and strong interactions. For this reason, they are all known as *gauge theories* (of essentially the Yang-Mills type). ([A1])

4.1. An interlude on differential field operators

To discuss the Maxwell differential equations of electromagnetism, we must first recall some concepts of differential 3-vector calculus. This calculus is an extension of normal differentiation applied to scalar and vector fields (see [S17] as a general reference for this section).

The basic differential operator is ∇ (called *del* or *nabla*) which is defined (in Cartesian coordinates) in terms of partial derivative operators as $\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial z)$. ∇ is used as a shorthand form to simplify expressions for the gradient, divergence, curl, and Laplacian.

- Gradient of a scalar field → 3-vector field

The *gradient* ∇s of a scalar field $x_\mu \to s(x_\mu)$ is defined as the 3-vector

$$\nabla s := (\partial s/\partial x, \partial s/\partial y, \partial s/\partial z).$$

The gradient vector ∇s always points in the direction of maximum change of the function $(x, y, z) \to s(x, y, z)$ and its length indicates the rate of change of the function in this direction.

- Divergence of a 3-vector field → scalar field

The scalar product of ∇ and a 3-vector field $x_\mu \to \boldsymbol{F}(x_\mu) = (F_x, F_y, F_z)$ is known as the *divergence* of $\boldsymbol{F}$

$$\nabla\cdot\boldsymbol{F} = \text{div}\,\boldsymbol{F} := \partial F_x/\partial x + \partial F_y/\partial y + \partial F_z/\partial z.$$

It is a scalar quantity which indicates the tendency of the field at a specific location to spread out. In technical terms, the divergence represents the volume density of the outward flux of a vector field from an infinitesimal volume around a given point. In physical terms, the divergence is the extent to which the vector field flux behaves like a source at a given location. A point at which the flux is outgoing has positive divergence, and is often called a *source* of the field. A point at which the flux is directed inward has negative divergence, and is often called a *sink* of the field ([W11]). If the divergence is zero at a given point then the flux that goes into the point must come out of it, i.e. this point is neither a source nor a sink of the field.

- Curl of a 3-vector field → 3-vector field

The vector product of ∇ and a 3-vector field $x_\mu \to \boldsymbol{F}(x_\mu)$ gives the *curl* of $\boldsymbol{F}$

$$\nabla\times\boldsymbol{F} = \text{curl}\,\boldsymbol{F} = (\partial F_z/\partial y - \partial F_y/\partial z,\ \ \partial F_x/\partial z - \partial F_z/\partial x,\ \ \partial F_y/\partial x - \partial F_x/\partial y).$$

Consequently, the curl of a 3-vector field is again a 3-vector field. The curl indicates the rotational ability of the vector field at a given location. To see what curl means globally, imagine dropping a leaf into a fluid. As the leaf moves along with the fluid flow, the curl measures the tendency of the leaf to rotate. If the curl is zero, then the leaf does not rotate as it moves through the fluid. Note, however, that some points in the field can have zero curl while others have nonvanishing curl.

- Laplacian of a scalar [vector] field → scalar [vector] field

The *Laplacian* $\nabla^2 s$ of a (twice-differentiable) scalar field $s = s(x_\mu)$ is defined as the scalar

$$\nabla^2 s := \nabla\cdot(\nabla s) = (\nabla\cdot\nabla)s = \partial^2 s/\partial x^2 + \partial^2 s/\partial y^2 + \partial^2 s/\partial z^2.$$

Thus the Laplacian is the divergence of the gradient. The operator ∇^2 is also called *Laplace operator* and denoted by Δ. The value of $\nabla^2 s$ at a particular point tells us how the value of s at that point compares to the average value of s at nearby surrounding points.

When the Laplace operator is applied to a vector field $x_\mu \to \boldsymbol{F}(x_\mu)$, it generates a vector field

$$\nabla^2 \boldsymbol{F} := \nabla(\nabla \cdot \boldsymbol{F}) - \nabla \times (\nabla \times \boldsymbol{F}).$$

In Cartesian coordinates, the resulting vector field is equal to the vector field of the scalar Laplacian applied to each vector component ([W11])

$$\nabla^2 \boldsymbol{F} = (\nabla^2 F_x, \nabla^2 F_y, \nabla^2 F_z).$$

4.2. Maxwell's equations

The basic laws of classical (i.e. non-quantum) electromagnetism are governed by the *Maxwell's equations* ([F2], [F3]). These equations were first completely formulated by Maxwell ([71]) in *Treatise on Electricity and Magnetism* in 1873. He took a set of known experimental laws (Gauss' Laws, Faraday's Law, and Ampère's Law) and unified them into a set of four differential equations.

The most common description of the electromagnetic field uses two 3-vector fields called the *electric field* $x_\mu \to \boldsymbol{E}(x_\mu) = (E_x, E_y, E_z)$ and the *magnetic field* $x_\mu \to \boldsymbol{B}(x_\mu) = (B_x, B_y, B_z)$. Maxwell's equations describe how these fields propagate, interact, and how they are influenced by charges. Maxwell's equations read as follows

- *Gauss' Magnetic Law*

(M1) $$\nabla \cdot \boldsymbol{B} = 0$$

- *Faraday's Law of Induction*

(M2) $$\nabla \times \boldsymbol{E} + \partial \boldsymbol{B}/\partial t = 0$$

- *Gauss' Electric Law*

(M3) $$\nabla \cdot \boldsymbol{E} = \rho/\varepsilon_0$$

- *Ampère-Maxwell's Law*

(M4) $$c^2 \nabla \times \boldsymbol{B} - \partial \boldsymbol{E}/\partial t = \boldsymbol{j}/\varepsilon_0,$$

where ρ is the *charge density*, ε_0 is the *electric constant*, and $\boldsymbol{j}$ is the *current density.* In general, the quantities ρ and $\boldsymbol{j}$ depend on time and position. Charge density ρ is the amount of charge per unit volume. It is a scalar. Current density 3-vector $\boldsymbol{j} = (j_x, j_y, j_z)$ is the current per unit area. It represents the flow of charge, e.g. the flow of electrons through a wire. The constant c^2 in (M4) is the square of the velocity of light. It appears because magnetism is in reality a relativistic effect of electricity. The constant ε_0 has been stuck in to make the units of electric current come out in a convenient way.

Equations (M3) and (M4), which have sources on the right-hand side, are called the *Maxwell Field Equations.* Equations (M1) and (M2) are called *Bianchi Identities*.

How many Maxwell's equations are altogether? Actually there are eight, although in vector notation we see only four: two 3-vector equations (M2) and (M4), and two scalar equations (M1) and (M3). But equations (M2) and (M4) have three components apiece. For example, equation (M2) amounts to

$$\partial E_z/\partial y - \partial E_y/\partial z - \partial B_x/\partial t = 0, \;\; \partial E_x/\partial z - \partial E_z/\partial x - \partial B_y/\partial t = 0, \;\; \partial E_y/\partial x - \partial E_x/\partial y - \partial B_z/\partial t = 0.$$

What do Maxwell's equations mean? The first Maxwell's equation $\nabla \cdot \boldsymbol{B} = 0$ says that there cannot be magnetic charges. While we have electric charges, there is no configuration with magnetic field vectors diverging from a single point – a magnetic monopole. Equation (M1) states that the magnetic field tends to wrap around things – since the divergence is zero, the fields tend to form closed loops.

The second equation (M2), that the curl $\nabla \times \boldsymbol{E}$ of $\boldsymbol{E}$ is $-\partial \boldsymbol{B}/\partial t$, is Faraday's law and is generally true. It tells us that a magnetic field that is changing in time will give rise to a

[71] James Clerk Maxwell (1831 – 1879) was a Scottish mathematician and scientist responsible for the classical theory of electromagnetic radiation.

circulating electric field. This equation describes many phenomena of great practical interest, such as those that occur in electric generators and transformers. A moving magnet must make an electric field. How that happens is said quantitatively by (M2).

The third equation (M3) − that the divergence $\nabla\cdot\boldsymbol{E}$ of $\boldsymbol{E}$ is the charge density ρ over ε_0 − is true in general. In dynamic as well as in static fields, this Gauss' law is always valid. The flux of $\boldsymbol{E}$ through any closed surface is proportional to the charge inside. Thus Gauss' law says that electric field lines diverge away from electric charges. More precisely, positive charges act as a source, whereas negative charges act as a sink of electric fields (cf. Section 4.1). Consequently, electric field lines begin and end only at charges or at infinity.

The last equation (M4) tells us that a flowing electric current gives rise to a magnetic field that circles the wire. In addition to this, it also says that an electric field that is changing in time gives rise to a magnetic field that encircles the electric field.

Equation (M4) was discovered by Maxwell. Before Maxwell's work the Ampère's law for steady current was known only as: $c^2\nabla\times\boldsymbol{B} = \boldsymbol{j}/\varepsilon_0$. Maxwell noticed that there was something strange about this equation. If one takes the divergence of this equation, the left-hand side will be zero, because the divergence of a curl $\nabla\cdot(\nabla\times\boldsymbol{B})$ is always zero.

Indeed, for any 3-vector field $\boldsymbol{A} = (A_x, A_y, A_z)$ we have

$$\nabla\cdot(\nabla\times\boldsymbol{A}) = [\partial/\partial x, \partial/\partial y, \partial/\partial z]\begin{bmatrix}\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}\\ \frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\\ \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\end{bmatrix} =$$

$$\partial/\partial x(\partial A_z/\partial y - \partial A_y/\partial z) + \partial/\partial y(\partial A_x/\partial z - \partial A_z/\partial x) + \partial/\partial z(\partial A_y/\partial x - \partial A_x/\partial y) =$$
$$(\partial^2 A_z/\partial x\partial y - \partial^2 A_z/\partial y\partial x) + (-\partial^2 A_y/\partial x\partial z + \partial^2 A_y/\partial z\partial x) + (\partial^2 A_x/\partial y\partial z - \partial^2 A_x/\partial z\partial y) = 0.$$

So the Ampère's law requires that the divergence $\nabla\cdot\boldsymbol{j}$ of $\boldsymbol{j}$ also be zero. But if the divergence of $\boldsymbol{j}$ is zero, then the total flux of current out of any closed surface is also zero. This can only be true in situations where the charge density is constant in time. The flux of current $\nabla\cdot\boldsymbol{j}$ from a closed surface is the decrease of the charge inside the surface. This certainly cannot in general be zero because we know that the charges can be moved from one place to another. For the general case, Maxwell modified Ampère's law to be read as (M4). Maxwell's addendum is the term $\partial\boldsymbol{E}/\partial t$, known as the *displacement current*. The presence of this term makes a great difference, for now there is the possibility for wave motion.

The *electromagnetic force*, also called the *Lorentz force*, explains how both moving and stationary charged particles interact. It is called the electromagnetic force because it includes the electrical force and the magnetic force.

The electrical force, like a gravitational force, decreases inversely as the square of the distance between charges. This relationship is called *Coulomb's law*. But it is not precisely true when charges are moving − the electrical forces depend also on the motions of the charges in a complicated way. One part of the force between moving charges we call the magnetic force. It is really one aspect of an electrical effect. Electrical and magnetic forces are closely related to each other. An electrical force in one frame of reference becomes a magnetic force in another frame, and vice versa. In other words, electrical and magnetic forces transform into each other under Lorentz transformation. That is why we call the subject *electromagnetism*.

There is an important general principle that makes it possible to treat electromagnetic forces in a relatively simple way. The force that acts on a particular charge − no matter how many other charges there are or how they are moving − depends only on the position of that particular charge, on the velocity of the charge, and on the amount of charge. We can write the electromagnetic force $\boldsymbol{F}$ on a charge q moving with a velocity $\boldsymbol{v}$ as

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v}\times\boldsymbol{B}). \tag{4.1}$$

The two terms represent the electrical force $q\boldsymbol{E}$ and the magnetic force $q(\boldsymbol{v}\times\boldsymbol{B})$, which are proportional to the charge q. The electrical force is independent of the particle's velocity. The magnetic force is more complicated and involves the vector cross product. The reason for it is that the direction of the magnetic force depends not only on the direction of the magnetic field but varies also according to the direction of the velocity vector. The expression $\boldsymbol{v}\times\boldsymbol{B}$ says that the magnetic force is perpendicular to both the magnetic field $\boldsymbol{B}$ and the velocity $\boldsymbol{v}$.

You can experience the first term $q\boldsymbol{E}$ of equation (4.1) when combing your hair. The second part can be demonstrated by passing a current through a wire which hangs above a bar magnet ([F2; Fig. 1-6]). The wire will move when a current is turned on because of the force $q(\boldsymbol{v}\times\boldsymbol{B})$. When a current exists, the charges inside the wire are moving, so they have a velocity $\boldsymbol{v}$, and the magnetic field from the magnet exerts a force on them, which results in pushing the wire sideways.

Equations (M1) through (M4), together with the Lorentz force equation (4.1), are all the laws of electrodynamics. Electromagnetic fields control (charged) particles through the Lorenz force, while charges control fields through Maxwell's equations. L. Susskind ([72]) paraphrasing of John Wheeler's slogan ([73]), puts this in following words: *fields tell charges how to move; charges tell fields how to vary* ([S17]).

From (M3) and (M4) we infer the equation (*the continuity equation for electric charge*)

$$\nabla\cdot\boldsymbol{j} = -\partial\rho/\partial t \tag{4.2}$$

which is at heart of electrodynamics and expresses the very fundamental law that electric charge is conserved — any flow of charge must come from some supply.

Indeed, by (M4) we have

$$0 = \nabla\cdot(c^2\nabla\times\boldsymbol{B}) = \nabla\cdot(\boldsymbol{j}/\varepsilon_0 + \partial\boldsymbol{E}/\partial t) = \nabla\cdot\boldsymbol{j}/\varepsilon_0 + \nabla\cdot(\partial\boldsymbol{E}/\partial t) = \nabla\cdot\boldsymbol{j}/\varepsilon_0 + \partial/\partial t(\nabla\cdot\boldsymbol{E}).$$

Thus applying (M3) we get $\nabla\cdot\boldsymbol{j} = -\partial\rho/\partial t$.

It follows from the continuity equation (4.2) that the rate of decrease of charge in any arbitrary volume V is due precisely and only to the flow of the charge through the walls of its surface; that is, no net charge can be created or destroyed in V. Since V can be made as small as we please, this means that *electric charge must be locally conserved:* a process in which charge is created at one point and destroyed at a distant one is not allowed, despite the fact that it conserves the total charge overall or 'globally'.

The ultimate reason for this is that the global form of charge conservation would necessitate the instantaneous propagation of signals, and this conflicts with special relativity ([74]).

---

[72] Leonard Susskind (1940 –) is an American physicist, a professor of theoretical physics at Stanford University.

[73] J. Wheeler summed up general relativity in words as follows: *"spacetime tells matter how to move; matter tells spacetime how to curve."* John Archibald Wheeler (1911 – 2008) was a prominent American theoretical physicist. He coined the term 'black hole'. Two of his students, Richard Feynman and Kip Thorne, received Nobel Prizes.

[74] **Question**: *Would you distinguish local conservation laws from global conservation laws.*
Feynman: *If a cat were to disappear in Pasadena and at the same time appear in Erice, that would be an example of global conservation of cats. This is not the way cats are conserved. Cats or charge or baryons are conserved in a much more continuous way. If any of these quantities begin to disappear in a region, then they begin to appear in a neighbouring region. Consequently, we can identify the flow of charge out of a region with the disappearance of charge inside the region. This identification of the divergence of a flux with the time rate of change of a charge density is called a local conservation law. A local conservation law implies that the total charge is conserved globally, but the reverse does not hold. However, relativistically it is clear that non-local global conservation laws cannot exist, since to a moving observer the cat will appear in Erice before it disappears in Pasadena.*

– From the question-and-answer session following a lecture by R. P. Feynman at the 1964 International School of Physics 'Ettore Majorana' in Erice [F7].

The most remarkable consequence of Maxwell's equations is that the combination of (M2) and (M4) contains the explanation of the radiation of electromagnetic effects over large distances. The reason is roughly something like this: suppose that somewhere we have a magnetic field which is increasing because, say, a current is turned on suddenly in a wire. Then by (M2) there must be a circulation of an electric field. As the electric field builds up to produce its circulation, then according to (M4) a magnetic circulation will be generated. But the building up of *this* magnetic field will produce a new circulation of the electric field, and so on. In this way, fields work their way through space without the need for charges or currents except at their source ([F2]).

Any fundamental theory in physics is expected to be in agreement with the principle of relativity, i.e. its equations have to be Lorentz covariant. It means that the equations of the theory remain true after a Lorentz transformation. A Yang-Mills theory is expected not only to be Lorentz covariant but also gauge invariant: its equations should remain unchanged under a gauge transformation. It should be noted that, in contrast to the principle of Lorentz covariance, the definition and meaning of gauge invariance are dependent on the specific theory under consideration.

Since electrodynamics is a simple example of a Yang-Mills theory we can use it to discuss the meaning of 'Lorentz covariance' and 'gauge invariance'. Let us start with Lorentz covariance ([75]).

4.3. Lorentz covariance of the Maxwell's equations

Lorentz covariance of Maxwell's equations is certainly the key link between classical electrodynamics and special relativity. Generally, it is demonstrated that Maxwell's equations are Lorentz covariant if and only if the electric and magnetic fields, and charge and current densities appearing in them transform according to Lorentz transformation laws. As is well known, this can be done basically in two ways: either transforming directly Maxwell's equations ('steep and difficult mountaineer's path') as Einstein originally did or employing the powerful and elegant tensorial approach in Minkowski spacetime ([R1]).

Maxwell's equations (M1) – (M4) are formulated by means of quantities of vector analysis in three dimensions. Now in the case of special relativity, time and space are inextricably mixed, and we must do the analogous things for four dimensions: instead of 3-vectors we have to use 4-vectors ([76]).

Let us start with the four-dimensional analogue ∇_μ of the differential operator $\boldsymbol{\nabla} = (\partial/\partial x, \partial/\partial y, \partial/\partial z)$ ([F5]). We might guess that the four-dimensional gradient should be $\nabla_\mu = (\partial/\partial t, \partial/\partial x, \partial/\partial y, \partial/\partial z)$. But this is wrong because this quantity does not behave as a 4-vector. The answer is that instead of the *incorrect* $(\partial/\partial t, \boldsymbol{\nabla})$, we must define the four-dimensional gradient operator by

$$\nabla_\mu := (\partial/\partial t, -\boldsymbol{\nabla}) = (\partial/\partial t, -\partial/\partial x, -\partial/\partial y, -\partial/\partial z).$$

Then the operator ∇_μ transforms in the same way as the position vector x_μ i.e. ∇_μ is a 4-vector (called also 4-*vector gradient* and sometimes denoted by ∂_μ). It simply means that the 4-gradient of a scalar is a 4-vector. Thus if ϕ is a Lorentz invariant scalar field then $\nabla_\mu\phi = (\partial\phi/\partial t, -\partial\phi/\partial x, -\partial\phi/\partial y, -\partial\phi/\partial z) = (\partial\phi/\partial t, -\boldsymbol{\nabla}\phi)$ is a 4-vector field.

The next thing that we have to discuss is the 4-dimensional analogue of the divergence of 3-vectors ([F5]). We *define* the *divergence* $\nabla_\mu a^\mu$ of the 4-vector $a_\mu = (a_t, \boldsymbol{a})$ as the scalar product (3.4) of ∇_μ and a_μ:

---

[75] Throughout the remainder of this article we shall generally (unless otherwise stated) use the natural units: $c = \hbar = 1$.

[76] As mentioned above, it is possible to verify Lorentz covariance starting from the original Maxwell equations (M1) – (M4). This involves however establishing the rather complicated transformation law for the fields $\boldsymbol{E}$ and $\boldsymbol{B}$. Thus the standard way in the literature is to reformulate Maxwell equations in terms of 4-vectors, Lorentz scalars and covariant rank-2 tensors. This allows to write Maxwell equations in a manifestly Lorentz covariant form.

(4.3) $$\nabla_\mu a^\mu := \nabla_\mu \eta^{\mu\nu} a_\nu = \partial a_t/\partial t - \partial(-a_x)/\partial x - \partial(-a_y)/\partial y - \partial(-a_z)/\partial z$$
$$= \partial a_t/\partial t + \partial a_x/\partial x + \partial a_y/\partial y + \partial a_z/\partial z = \partial a_t/\partial t + \boldsymbol{\nabla}\cdot\boldsymbol{a},$$

where $\boldsymbol{\nabla}\cdot\boldsymbol{a}$ is the ordinary divergence of the 3-vector $\boldsymbol{a}$ ([77]). The divergence $\nabla_\mu a^\mu$ is an invariant and gives the same answer in all coordinate systems which differ by a Lorentz transformation.

The last operator we want to consider ([78]) is the scalar product of the gradient operator ∇_μ with itself ([F5]). In three dimensions, such a product gives the Laplacian (see Section 4.1)

$$\nabla^2 = \boldsymbol{\nabla}\cdot\boldsymbol{\nabla} = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2.$$

In four dimensions we get by (3.4)

$$\nabla_\mu \nabla^\mu = \partial/\partial t \partial/\partial t - (-\partial/\partial x)(-\partial/\partial x) - (-\partial/\partial \mathrm{y})(-\partial/\partial \mathrm{y}) - (-\partial/\partial \mathrm{z})(-\partial/\partial \mathrm{z}) = \partial^2/\partial t^2 - \nabla^2.$$

This operator, which is the analogue of the three-dimensional Laplacian, is called the *d'Alembertian* and is frequently written as $\Box^2$ ([79])

(4.4) $$\Box^2 := \nabla_\mu \nabla^\mu = \partial^2/\partial t^2 - \nabla^2.$$

It is manifestly a Lorentz scalar operator since it is built from the contraction of indices (i.e. scalar product) on the two 4-gradient operators.

Now we are going to discuss the 3-vectors $\boldsymbol{j}$, $\boldsymbol{E}$ and $\boldsymbol{B}$. Let us start with the current density $\boldsymbol{j}$. It can be shown that the four quantities ρ, j_x, j_y, j_z transform as a four-vector ([80]). Thus we can write them as the 4-vector $j_\mu := (\rho, \boldsymbol{j})$ and call it the *electromagnetic 4-current density*. It represents the distribution of electric charges and currents in space and time.

The last step is to consider the 3-vectors $\boldsymbol{E}$ and $\boldsymbol{B}$. Since these 3-vectors describing the electric and magnetic fields have three components each, there is clearly no way in which they can be 'assembled' into 4-vectors. However, we may note that in four dimensions an antisymmetric rank-2 tensor has $(4\times3)/2 = 6$ independent components ([81]). It suggests that perhaps we could group the electric and magnetic fields together into a single antisymmetric rank-2 tensor $F_{\mu\nu} = F_{\mu\nu}(x_\mu)$ ([P7]). It turns out ([P7]) that we should define its components in terms of $\boldsymbol{E}$ and $\boldsymbol{B}$ as follows ([82]):

(4.5) $$F_{\mu\nu} := \begin{bmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & -B_z & B_y \\ E_y & B_z & 0 & -B_x \\ E_z & -B_y & B_x & 0 \end{bmatrix}$$

Rows of the matrix (4.5) are labelled by μ and columns by ν. For the matrix components, we will use the notation F_{00}, F_{01}, ..., F_{03}, F_{10}, F_{11}, ..., F_{13}, ..., F_{33} or F_{tt}, F_{tx}, ..., F_{tz}, F_{xt}, F_{xx}, ..., F_{xz}, ..., F_{zz}.

---

[77] Note that one has to be careful with the signs. Some of the minus signs come from the definition of the scalar product (3.3); the others are required because the space components of ∇_μ are $-\partial/\partial x$, etc.

[78] We do not yet have the equivalents of the cross product and the curl operator – we will get to them later on.

[79] There are a variety of notations for the d'Alembertian. The most common are the box symbol $\Box$ and the box-squared symbol $\Box^2$. Another way to write the d'Alembertian is ∂^2. Some people define the d'Alembertian with the opposite sign to (4.4), so you will have to be careful when reading the literature.

[80] Roughly speaking, an electric charge is Lorentz invariant since all observers will agree on the number of electrons in a given closed spatial region, and so they will agree on the amount of charge. Thus ρ must transform like the 0 component of a 4-vector. In the same way, we may consider the spatial components j_i of j_μ. which are related to a flow of charge.

[81] If a 2-rank tensor F is represented by a 4×4 matrix $F_{\mu\nu}$ then F is *antisymmetric* (or *skew-symmetric*) if $F_{\mu\nu} = -F_{\nu\mu}$, i.e. its transpose $(F_{\mu\nu})^T$ is equal to its negative. So, out of sixteen components of $F_{\mu\nu}$ we get only six different objects.

[82] We shall justify this a little later.

The 'upstairs' version of $F_{\mu\nu}$ is ([83])

(4.6) $$F^{\mu\nu} = \begin{bmatrix} 0 & E_x & E_y & E_z \\ -E_x & 0 & -B_z & B_y \\ -E_y & B_z & 0 & -B_x \\ -E_z & -B_y & B_x & 0 \end{bmatrix}$$

Now we are ready to see how the Maxwell's equations look when expressed in terms of $F_{\mu\nu}$, $F^{\mu\nu}$ and j_μ. Maxwell's field equations (M3) and (M4) become ([84]):

(M3+M4) $$\nabla^\mu F_{\mu\nu} = j_\nu/\varepsilon_0,$$

where $j_\nu = (\rho, \boldsymbol{j})$ ([P7]).

Indeed, the vector $\nabla^\mu F_{\mu\nu}$ has the components $\nabla^\mu F_{\mu t}$, $\nabla^\mu F_{\mu x}$, $\nabla^\mu F_{\mu y}$, $\nabla^\mu F_{\mu z}$ so we can write (M3+M4) in the form

$$\nabla^\mu F_{\mu\nu} = \begin{bmatrix} \nabla^\mu F_{\mu t} \\ \nabla^\mu F_{\mu x} \\ \nabla^\mu F_{\mu y} \\ \nabla^\mu F_{\mu z} \end{bmatrix} = \begin{bmatrix} \rho/\varepsilon_0 \\ j_x/\varepsilon_0 \\ j_y/\varepsilon_0 \\ j_z/\varepsilon_0 \end{bmatrix}.$$

Now let us show that

$$\nabla^\mu F_{\mu\nu} = [\,\partial/\partial t \quad \nabla] \begin{bmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & -B_z & B_y \\ E_y & B_z & 0 & -B_x \\ E_z & -B_y & B_x & 0 \end{bmatrix} = \begin{bmatrix} \nabla\cdot\boldsymbol{E} \\ -\frac{\partial \boldsymbol{E}}{\partial t} + \nabla \times \boldsymbol{B} \end{bmatrix}.$$ ([85])

For $\nu = t$ we have to multiply $[\,\partial/\partial t \quad \nabla] = [\,\partial/\partial t \quad \partial/\partial x \quad \partial/\partial y \quad \partial/\partial z]$ by the first column of $F_{\mu\nu}$

$$\nabla^\mu F_{\mu t} = \partial 0/\partial t + \partial E_x/\partial x + \partial E_y/\partial y + \partial E_z/\partial z = \nabla\cdot\boldsymbol{E}.$$

But by (M3+M4) $\nabla^\mu F_{\mu t} = \rho/\varepsilon_0$ and this implies that $\nabla\cdot\boldsymbol{E} = \rho/\varepsilon_0$. Thus the field equation (M3+ M4) embodies (M3). Now for $\nu = x$ we get

$$\begin{aligned} \nabla^\mu F_{\mu x} &= \partial(-E_x)/\partial t + \partial 0/\partial x + \partial B_z/\partial y + \partial(-B_y)/\partial z \\ &= -\partial E_x/\partial t + (\partial B_z/\partial y - \partial B_y/\partial z) = -\partial E_x/\partial t + (\nabla\times\boldsymbol{B})_x. \end{aligned}$$

But $\nabla^\mu F_{\mu x} = j_x/\varepsilon_0$ and consequently $(\nabla\times\boldsymbol{B})_x = j_x/\varepsilon_0 + \partial E_x/\partial t$.

Similarly we can show that $(\nabla\times\boldsymbol{B})_y = j_y/\varepsilon_0 + \partial E_y/\partial t$ and $(\nabla\times\boldsymbol{B})_z = j_z/\varepsilon_0 + \partial E_z/\partial t$. Putting these three equalities together we get the equation (M4): $\nabla\times\boldsymbol{B} = \boldsymbol{j}/\varepsilon_0 + \partial\boldsymbol{E}/\partial t$.

Consequently (M3+M4) is equivalent to Maxwell's field equations (M3) and (M4).

The remaining two Maxwell's equations (M1) and (M2) become ([86]) ([P7]):

(M1+M2) $$\nabla_\mu F^{\nu\rho} + \nabla_\nu F^{\rho\mu} + \nabla_\rho F^{\mu\nu} = 0.$$

---

[83] Recall that index raising (and lowering) is defined through the Minkowski metric $\eta^{\mu\nu}$. Now, if you want to raise two indices you need to apply it twice: $F^{\mu\nu} = \eta^{\mu\rho} F_{\rho\sigma} \eta^{\sigma\nu}$. In a more formal language lowering and raising indices is a way to construct isomorphisms between covariant and contravariant tensorial spaces. We use the metric tensor because it helps to map basis vectors e_i to dual basis vectors.

[84] Reminder: $\nabla^\mu := \eta^{\nu\mu}\nabla_\mu = (\partial/\partial t, \nabla)$, where $\eta^{\nu\mu}$ is the Minkowski metric.

[85] Notice that the vector $-\frac{\partial \boldsymbol{E}}{\partial t} + \nabla \times \boldsymbol{B}$ has three components: $-\partial E_i/\partial t + (\nabla \times \boldsymbol{B})_i, i = x, y, z$.

[86] In contrast to equation (M1+M4) for $F_{\mu\nu}$ which couples to the electrical current density j_μ this equation has a zero right-hand side. This is a consequence of the incomplete symmetry between electric and magnetic field, arising from the nonexistence of magnetic monopoles ('magnetic charges') ([B4]).

The indices μ, ν, and ρ can take on any of the four values (0, 1, 2, 3) or (t, x, y, z). No matter which of these values we assign to μ, ν, and ρ, the above equation gives a result of zero.

To see this, let us compute the components of the expression on the left side of equation (M1+M2). It follows from the antisymmetry of $F^{\mu\nu}$ that this expression is totally antisymmetric in (μνρ) (i.e. it changes sign under any exchange of a pair of indices).

Moreover, for repeating indices the corresponding components are 0. For example, taking (μ, ν, ρ) = (0, 0, 0) or (μ, ν, ρ) = (1, 0, 0) and applying (M1+M2) we get

$$0 = \partial F^{00}/\partial t + \partial F^{00}/\partial t + \partial F^{00}/\partial t = \partial 0/\partial t + \partial 0/\partial t + \partial 0/\partial t = 0,$$

$$0 = \partial F^{00}/\partial x + \partial F^{01}/\partial t + \partial F^{10}/\partial t = \partial 0/\partial t + \partial E_x/\partial t - \partial E_x/\partial t = 0.$$

Consequently, it is enough to compute the components for the following four assignments of indices only: (0, 1, 2), (0, 1, 3), (0, 2, 3) and (1, 2, 3). Applying (M1+M2) we get then the following equations

$$\begin{aligned} 0 &= \partial F^{12}/\partial t + \partial F^{20}/\partial x + \partial F^{01}/\partial y = -\partial B_z/\partial t - \partial E_y/\partial x + \partial E_x/\partial y \\ &= -\partial B_z/\partial t - (\boldsymbol{\nabla}\times\boldsymbol{E})_z \end{aligned}$$

$$\begin{aligned} 0 &= \partial F^{13}/\partial t + \partial F^{30}/\partial x + \partial F^{01}/\partial z = \partial B_y/\partial t - \partial E_z/\partial x + \partial E_x/\partial z \\ &= \partial B_y/\partial t + (\boldsymbol{\nabla}\times\boldsymbol{E})_y \end{aligned}$$

$$\begin{aligned} 0 &= \partial F^{23}/\partial t + \partial F^{30}/\partial y + \partial F^{02}/\partial z = -\partial B_x/\partial t - \partial E_z/\partial y + \partial E_y/\partial z \\ &= -\partial B_x/\partial t - (\boldsymbol{\nabla}\times\boldsymbol{E})_x \end{aligned}$$

Putting these three equalities together we get (M2): $\boldsymbol{\nabla}\times\boldsymbol{E} = -\partial\boldsymbol{B}/\partial t$. Now for (μ, ν, ρ) = (1, 2, 3) we obtain equation (M1):

$$0 = \partial F^{23}/\partial x + \partial F^{31}/\partial y + \partial F^{12}/\partial z = -\partial B_x/\partial x - \partial B_y/\partial y - \partial B_z/\partial z = -\boldsymbol{\nabla}\cdot\boldsymbol{B}.$$

The tensor $F_{\mu\nu}$ is called the *electromagnetic field tensor* (or the *field-strength tensor*). It combines the magnetic and the electric field into a single object. Its source is the electric 4-current j_μ – see (M3+M4). We will show in the next section that $F_{\mu\nu}$ can be written as

$$F_{\mu\nu} = \nabla_\mu A_\nu - \nabla_\nu A_\mu \tag{4.7}$$

for some 4-vector $A_\mu = A_\mu(x_\mu)$. The continuity equation (4.2) can be expressed in the following form $\nabla^\mu j_\mu = 0$. The field equation (M3+M4) $\nabla^\mu F_{\mu\nu} = j_\nu/\varepsilon_0$ then of course automatically embodies (4.2). The mathematical reason it does so is that $F_{\mu\nu}$ is a four-dimensional kind of 'curl' ([A1]).

In three dimensions the transformation properties of the curl $\boldsymbol{\nabla}\times\boldsymbol{a}$ are the same as the transformation properties of two 3-vectors − the 3-vector $\boldsymbol{a}$ and the gradient operator $\boldsymbol{\nabla}$ which also behaves like a 3-vector. Our electromagnetic quantity $F_{\mu\nu}$ is a tensor of the second rank in four dimensions. It transforms however in a special way which we will see in a moment − it is just the way a product of 3-vectors transforms.

We are going to show that Maxwell's equations are Lorentz covariant. But since the equations are now formulated using the 2-rank tensor $F_{\mu\nu}$, the question is, what does 'covariant tensor' mean. We already know that a vector with four components is Lorentz covariant if it transforms as the position vector. But $F_{\mu\nu}$ is not a vector so we have to clarify the meaning of the covariance in this case ([A1]).

Generally speaking, a physical quantity is said to be Lorentz *covariant* if it transforms under a given representation of the Lorentz group ([87]). For our purposes, it means that under a Lorentz

[87] According to the representation theory of the Lorentz group, Lorentz covariant quantities are built out of scalars, four-vectors, four-tensors, and spinors − see [S3] for details.

transformation, the components of $F_{\mu\nu}$ will transform into definite linear combinations of themselves. And this is easy to show ([F6]).

Indeed, let us recall the Lorentz transformation formulas (3.5) for two Lorentz frames L and L‘ (note: c = 1):

(4.8) $$t' = \gamma(t - \beta x), \quad x' = \gamma(x - \beta t), \quad y' = y, \quad z' = z,$$

where $\beta = v$ and $\gamma = 1/\sqrt{1 - \beta^2}$, and consider the general antisymmetric vector combination

$$c_{\mu\nu} = a_\mu b_\nu - a_\nu b_\mu,$$

where a_μ and b_μ are 4-vectors. Then a_μ and b_μ transform according to (4.8). Now let us transform the components of $c_{\mu\nu}$. We start with c_{tx}:

$$c'_{tx} = a'_t b'_x - a'_x b'_t = \gamma(a_t - \beta a_x)\gamma(b_x - \beta b_t) - \gamma(a_x - \beta a_t)\gamma(b_t - \beta b_x) = \gamma^2[(a_t b_x - a_x b_t) - \beta^2(a_t b_x - a_x b_t)] = \gamma^2(1 - \beta^2)(a_t b_x - a_x b_t) = a_t b_x - a_x b_t = c_{tx}.$$

Let us do one more

$$c'_{ty} = a'_t b'_y - a'_y b'_t = \gamma(a_t - \beta a_x)b_y - a_y\gamma\,(b_t - \beta b_x) = \gamma[(a_t b_y - a_y b_t) - \beta(a_x b_y - a_y b_x)] = \gamma(c_{ty} - \beta c_{xy}).$$

And in the same way we get

$$c'_{tz} = \gamma(c_{tz} - \beta c_{xz}), \qquad c'_{xy} = \gamma(c_{xy} - \beta c_{ty}), \; c'_{yz} = c_{yz}, \qquad c'_{zx} = \gamma(c_{zx} - \beta c_{zt}).$$

Of course, $c'_{\mu\nu} = -c'_{\nu\mu}$ and $c'_{\mu\mu} = 0$.

Replacing $c_{\mu\nu}$ by $F_{\mu\nu}$ we infer that the components of $F_{\mu\nu}$ transform into linear combinations of themselves. Hence the tensor $F_{\mu\nu}$ is Lorentz covariant.

We now state a very useful and important fact ([A1]). Suppose we ‘dot’ an upstairs 4-vector a^μ into a covariant second-rank tensor $B_{\mu\nu}$, via the operation $a^\mu B_{\mu\nu}$, where as always a sum on the repeated index μ is understood. Then this quantity transforms as a 4-vector, via its ‘loose’ index ν. An example is provided by the quantity $\nabla^\mu F_{\mu\nu}$ which enters on the left-hand side of the Maxwell’s equations in the form (M3+M4). Since j_ν/ε_0 is also a 4-vector we infer that both sides of the equation (M3+M4) transform as a 4-vector. Consequently, the equation (M3+M4) is Lorentz covariant. In a similar way, we can show that the equation (M1+M2) is also Lorentz covariant. It implies that the original Maxwell’s equations (M1) – M(4) are Lorentz covariant ([A1]).

Finally, we calculate the following Lorentz invariant from the electromagnetic tensor which we will need later when discussing the Lagrangian formalism

(4.9) $$F_{\mu\nu}F^{\mu\nu} = 2(\boldsymbol{B}^2 - \boldsymbol{E}^2).$$

Mind that $F_{\mu\nu}F^{\mu\nu}$ is **not** a matrix being equal to the product of the matrices $F_{\mu\nu}$ and $F^{\mu\nu}$. It is a scalar (reminder: implicit summation on repeated indices)

$$\begin{aligned} F_{\mu\nu}F^{\mu\nu} &:= \textstyle\sum_\mu \sum_\nu F_{\mu\nu}F^{\mu\nu} = \sum_\nu F_{t\nu}F^{t\nu} + \sum_\nu F_{x\nu}F^{x\nu} + \sum_\nu F_{y\nu}F^{y\nu} + \sum_\nu F_{z\nu}F^{z\nu} \\ &= (0 - E_x^2 - E_y^2 - E_z^2) + (-E_x^2 + 0 + B_z^2 + B_y^2) + \\ &\quad (-E_y^2 + B_z^2 + 0 + B_x^2) + (-E_z^2 + B_y^2 + B_x^2 + 0) \\ &= 2(B_x^2 + B_y^2 + B_z^2) - 2(E_x^2 + E_y^2 + E_z^2) \\ &= 2(\boldsymbol{B}^2 - \boldsymbol{E}^2). \end{aligned}$$

4.4. Gauge invariance of Maxwell's equations

In classical electromagnetism, and especially in quantum mechanics, it is convenient to introduce the electromagnetic 4-vector potential $A_\mu(x)$ in place of the fields $\boldsymbol{E}$ and $\boldsymbol{B}$ ([88]).

We have already remarked that Maxwell's equations (M1) and (M2) are known as Bianchi identities. They are not field equations, since there are no sources; rather, they impose constraints on the electric and magnetic fields. We want to write them in terms of the potential A_μ ([F3]).

We begin with (M1), i.e. $\nabla\cdot\boldsymbol{B} = 0$ – the simplest of the equations. It is known that if the divergence $\nabla\cdot\boldsymbol{a}$ of a 3-vector $\boldsymbol{a}$ is 0 then it implies that $\boldsymbol{a}$ is the curl of something ([89]). So, if we write

(4.10) $$\boldsymbol{B} = \nabla\times\boldsymbol{A},$$

which defines the magnetic 3-vector potential $\boldsymbol{A} = \boldsymbol{A}(t, \boldsymbol{x})$, we have already solved one of Maxwell's equations ([90]): $\nabla\cdot\boldsymbol{B} = \nabla\cdot(\nabla\times\boldsymbol{A}) = 0$.

We take next the equation (M2), $\nabla\times\boldsymbol{E} = -\partial\boldsymbol{B}/\partial t$. If we write $\boldsymbol{B}$ as $\nabla\times\boldsymbol{A}$ and differentiate with respect to t, we can write Faraday's law (M2) in the form $\nabla\times\boldsymbol{E} = -\partial/\partial t\,(\nabla\times\boldsymbol{A})$. Hence $\nabla\times\boldsymbol{E} = -\nabla\times(\partial\boldsymbol{A}/\partial t)$ and consequently $\nabla\times(\boldsymbol{E} + \partial\boldsymbol{A}/\partial t) = 0$. We see that $\boldsymbol{E} + \partial\boldsymbol{A}/\partial t$ is a 3-vector whose curl is equal to zero. Therefore that vector is the gradient of something:

$$\boldsymbol{E} + \partial\boldsymbol{A}/\partial t = -\nabla V,$$

which defines the electric scalar potential $V = V(t, \boldsymbol{x})$ (the minus for technical convenience). So the equation (M2) can be put in the form

(4.11) $$\boldsymbol{E} = -\partial\boldsymbol{A}/\partial t - \nabla V.$$

We have solved two of Maxwell's equations already, and we have found that to describe the electromagnetic fields $\boldsymbol{E}$ and $\boldsymbol{B}$, we need four potential functions: a scalar potential V and a 3-vector potential $\boldsymbol{A}$, which is, of course, three functions $\boldsymbol{A} = (A_x, A_y, A_z)$.

A magnetic field exerts a force on a current, and a current density $\boldsymbol{j}$ has interaction energy given by $-\boldsymbol{j}\cdot\boldsymbol{A}$. This is an important formula that establishes the coupling of charged particles to the magnetic field through the vector potential. As will be shown below, $\boldsymbol{A}$ is, in fact, the spatial component of the 'gauge field' A_μ, a concept which forms the primary focus of this paper.

The origin of gauge invariance in classical electromagnetism lies in the fact that the potentials $\boldsymbol{A}$ and V are not unique for given physical fields $\boldsymbol{E}$ and $\boldsymbol{B}$. The transformations that $\boldsymbol{A}$ and V may undergo while preserving $\boldsymbol{E}$ and $\boldsymbol{B}$ (and hence Maxwell's equations) unchanged are called *gauge transformations*, and the associated invariance of Maxwell's equations is called *gauge invariance*.

What are these transformations? Clearly, $\boldsymbol{A}$ can be changed by

(4.12) $$\boldsymbol{A} \to \boldsymbol{A}' = \boldsymbol{A} + \nabla\Lambda,$$

where $\Lambda = \Lambda(t, \boldsymbol{x})$ is an arbitrary ([91]) scalar (real) function of position and time, with no change in $\boldsymbol{B}$ since $\nabla\times\nabla f = \boldsymbol{0}$, for any scalar function f with continuous mixed partial derivatives ([92]). To preserve $\boldsymbol{E}$, V must then change simultaneously by

---

[88] The introduction of the electromagnetic potential A_μ is motivated by the fact that two observers do not necessarily agree whether or not there is a non-zero magnetic field present in the system. In order to develop a consistent description that is valid for all observers, the electric and magnetic field components must be allowed to get mixed through coordinate transformations. The first step to do this is to introduce the 4-vector A_μ which combines the electric and magnetic potentials into a single object ([S2]). This is exactly this 4-vector which we announced in (4.7).

[89] This is so called Poincaré Lemma for anti-curl.

[90] Reminder: the divergence of a curl $\nabla\cdot(\nabla\times\boldsymbol{A})$ is always zero - see Section 4.2.

(4.13) $$V \to V' = V - \partial\Lambda/\partial t.$$

Then neither $\boldsymbol{B}$ nor $\boldsymbol{E}$ obtained from (4.10) and (4.11) is changed.

To summarise, if a given set of electric and magnetic fields $\boldsymbol{E}$ and $\boldsymbol{B}$ is described by a scalar potential V and 3-vector potential $\boldsymbol{A}$ according to (4.10) and (4.11), then the identical physical situation (i.e. identical electric and magnetic fields) is equally well described by a new pair of scalar and 3-vector potentials, related to the original pair by the gauge transformations given in (4.12) and (4.13), where Λ is an arbitrary function of position and time.

What is this gauge invariance (i.e. *gauge symmetry*) good for? We can in fact use the gauge invariance to our advantage, by making a convenient and simplifying gauge choice for the scalar and 3-vector potentials. We have one arbitrary function (i.e. $\Lambda(t, \boldsymbol{x})$) at our disposal, and so this allows us to impose one functional relation on the potentials V and $\boldsymbol{A}$. For our purposes, the most useful gauge choice is to use this freedom to impose the *Lorenz gauge condition* ([93])

(4.14) $$\boldsymbol{\nabla}\cdot\boldsymbol{A} = -\partial V/\partial t.$$ ([94])

Substituting (4.10) and (4.11) into Maxwell's field equations (M3) and (M4), and using the Lorenz gauge (4.14), we get ([95])

(4.15) $$\boldsymbol{\nabla}^2 V - \partial^2 V/\partial t^2 = -\rho/\varepsilon_0 \text{ and } \boldsymbol{\nabla}^2\boldsymbol{A} - \partial^2\boldsymbol{A}/\partial t^2 = -\boldsymbol{j}/\varepsilon_0.$$

These equations can be combined into a single compact equation by introducing the potential $A_\mu := (V, \boldsymbol{A})$ and using the d'Alembertian (4.4):

(4.16) $$\Box^2 A_\mu = j_\mu/\varepsilon_0.$$

The wave operator $\Box^2 := \nabla_\mu\nabla^\mu$ is manifestly a Lorentz operator, since it is built from the contraction of indices on the two gradient operators. Since we have already established that j_μ is a 4-vector, (4.16) therefore implies that A_μ is a 4-vector. It is called the *4-vector potential* (or *4-potential* for short). The ambiguity of A_μ up to a gauge transformation has led to the term *gauge field*. Equation (4.16) states that the current density j_μ is the source of the gauge field A_μ, and the field can propagate as a travelling wave with constant velocity c ([H13]).

The interaction energy density between matter and the electromagnetic field is the sum of electric and magnetic contributions, given by $\rho V - \boldsymbol{j}\cdot\boldsymbol{A}$. This can be expressed in the Lorentz invariant form: $j^\mu A_\mu$ ([H13]).

Note that the Lorenz gauge condition (4.14) translates in the four-dimensional language into

(4.17) $$\nabla^\mu A_\mu = 0,$$

which says that the divergence of the 4-potential A_μ is zero ([96]).

The energy of a relativistic charged particle moving in an electromagnetic field is given by

$$E = \sqrt{(\boldsymbol{p} - q\boldsymbol{A})^2 + m^2} + qV,$$

---

[91] More precisely, Λ is usually assumed to be smooth, i.e. infinitely differentiable scalar function.

[92] By the Clairaut's theorem on equality of mixed partials ([W11]), $\boldsymbol{\nabla}\times\boldsymbol{\nabla}f = (\frac{\partial}{\partial y}\frac{\partial f}{\partial z} - \frac{\partial}{\partial z}\frac{\partial f}{\partial y}, \frac{\partial}{\partial z}\frac{\partial f}{\partial x} - \frac{\partial}{\partial x}\frac{\partial f}{\partial z}, \frac{\partial}{\partial x}\frac{\partial f}{\partial y} - \frac{\partial}{\partial y}\frac{\partial f}{\partial x}) = (\frac{\partial}{\partial y}\frac{\partial f}{\partial z} - \frac{\partial}{\partial y}\frac{\partial f}{\partial z}, \frac{\partial}{\partial z}\frac{\partial f}{\partial x} - \frac{\partial}{\partial z}\frac{\partial f}{\partial x}, \frac{\partial}{\partial x}\frac{\partial f}{\partial y} - \frac{\partial}{\partial x}\frac{\partial f}{\partial y}) = (0, 0, 0) = \boldsymbol{0}$.

[93] Choosing the $\boldsymbol{\nabla}\cdot\boldsymbol{A}$ is called *choosing a gauge*. Equation (4.14) is called the *Lorenz gauge*.
Note that, contrary to the belief of many physicists, this gauge choice was introduced by the Danish physicist Ludvig Lorenz (1829 – 1891), and not the Dutch physicist Hendrik Lorentz (1853 – 1928) who is responsible for the Lorentz transformation. Adding to the confusion is that unlike many other gauge choices that one encounters, the Lorenz gauge is, as we shall see later, Lorentz invariant ([P7]).

[94] The function Λ is to be chosen to satisfy the equation $\boldsymbol{\nabla}^2\Lambda - \partial^2\Lambda/\partial t^2 = -\boldsymbol{\nabla}\cdot\boldsymbol{A} - \partial V/\partial t$.

[95] Reminder: $\boldsymbol{\nabla}^2\boldsymbol{A} := \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{A}) - \boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{A})$ – see Section 4.1.

[96] Equation (4.17) indicates that the gauge field A_μ itself does not have a charge.

where $\boldsymbol{p}$ is the momentum 3-vector (reminder: c = 1). The preceding equation can be written as

$$E - qV = \sqrt{(\boldsymbol{p} - q\boldsymbol{A})^2 + m^2}.$$

Comparing this to the energy of a free particle $E = \sqrt{\boldsymbol{p}^2 + m^2}$ we infer that the electromagnetic coupling appears through replacements $E \to E - qV$ and $\boldsymbol{p} \to \boldsymbol{p} - q\boldsymbol{A}$. But since $\mathrm{p}_\mu = (E, \boldsymbol{p})$ and $A_\mu = (V, \boldsymbol{A})$ it means the following 4-vector replacement ([H13])

$$\mathrm{p}_\mu \to \mathrm{p}_\mu - qA_\mu. \tag{4.18}$$

In the quantum form of the theory, this replacement corresponds to the substitution of the derivative ∇_μ by the covariant derivative D_μ (see Section 5.2).

Now let us note that our definition (4.5) is consistent with $A_\mu := (V, \boldsymbol{A})$ i.e., the 4-potential A_μ fulfils the equation (4.7): $F_{\mu\nu} = \nabla_\mu A_\nu - \nabla_\nu A_\mu$.

Indeed, put $G_{\mu\nu} := \nabla_\mu A_\nu - \nabla_\nu A_\mu$. Then

$G_{\alpha\alpha} = \nabla_\alpha A_\alpha - \nabla_\alpha A_\alpha = 0$

$G_{\mu\nu} = \nabla_\mu A_\nu - \nabla_\nu A_\mu = -(\nabla_\nu A_\mu - \nabla_\mu A_\nu) = -G_{\nu\mu}$

Applying (4.11) we obtain

$G_{tx} = \nabla_t A_x - \nabla_x A_t = \partial A_x/\partial \mathrm{t} + \partial A_t/\partial \mathrm{x} = \partial A_x/\partial \mathrm{t} + \partial V/\partial \mathrm{x} = -E_x = F_{tx}$

$G_{ty} = \nabla_t A_y - \nabla_y A_t = \partial A_y/\partial \mathrm{t} + \partial A_t/\partial \mathrm{y} = \partial A_y/\partial \mathrm{t} + \partial V/\partial \mathrm{y} = -E_y = F_{ty}$

$G_{tz} = \nabla_t A_z - \nabla_z A_t = \partial A_z/\partial \mathrm{t} + \partial A_t/\partial \mathrm{y} = \partial A_z/\partial \mathrm{t} + \partial V/\partial \mathrm{z} = -E_z = F_{tz}$

Now by (4.10) we get

$G_{xy} = \nabla_x A_y - \nabla_y A_x = -\partial A_y/\partial \mathrm{x} + \partial A_x/\partial \mathrm{y} = -(\boldsymbol{\nabla}\times\boldsymbol{A})_\mathrm{z} = -B_z = F_{xy}$

$G_{xz} = \nabla_x A_z - \nabla_z A_x = -\partial A_z/\partial \mathrm{x} + \partial A_x/\partial \mathrm{z} = (\boldsymbol{\nabla}\times\boldsymbol{A})_\mathrm{y} = B_y = F_{xz}$

$G_{yz} = \nabla_y A_z - \nabla_z A_y = -\partial A_z/\partial \mathrm{y} + \partial A_y/\partial \mathrm{z} = -(\boldsymbol{\nabla}\times\boldsymbol{A})_\mathrm{x} = -B_x = F_{xz}$

Consequently $F_{\mu\nu} = G_{\mu\nu} = \nabla_\mu A_\nu - \nabla_\nu A_\mu$.

The final step is to note that the gauge transformations (4.12) and (4.13) can be written in the form ([97]):

$$A_\mu \to A'_\mu = A_\mu - \nabla_\mu \Lambda. \tag{4.19}$$

Under the gauge transformation (4.19) the electromagnetic tensor $F_{\mu\nu}$ remains unchanged ([98])

$$F_{\mu\nu} \to F'_{\mu\nu} = F_{\mu\nu}.$$

Indeed,

$$F'_{\mu\nu} = \nabla_\mu A'_\nu - \nabla_\nu A'_\mu = \nabla_\mu(A_\nu - \nabla_\nu\Lambda) - \nabla_\nu(A_\mu - \nabla_\mu\Lambda)$$
$$= \nabla_\mu A_\nu - \nabla_\mu\nabla_\nu\Lambda - \nabla_\nu A_\mu + \nabla_\nu\nabla_\mu\Lambda = \nabla_\mu A_\nu - \nabla_\nu A_\mu = F_{\mu\nu}.$$

It means that $F_{\mu\nu}$ is gauge invariant and so, therefore, are Maxwell's field equations in the form (M3+M4).

Consequently, (4.16) describes the 'Lorentz-covariant and gauge-invariant field equations' satisfied by A_μ. ([99])

---

[97] Notice that the gauge transformation (4.19) does not act on spacetime but on the field space. It is thus an internal symmetry in the terminology of Section 3.5.

[98] It implies that the 4-vector potential A_μ is itself not physically observable since A_μ and A'_μ give rise to the same electromagnetic field $F_{\mu\nu}$.

[99] From a mathematical point of view the electromagnetic potential A_μ is a connection on the field space dictating how vector fields are going to displace when they move along specific paths along the base manifold $\mathbb{M}$. Gauge invariance allows for the establishment of a principal vector bundle structure, with different connections, that is, electromagnetic potentials, over each open set of the covering of the base space. In the overlapping zones of two open sets the respective connections must only differ by a gauge transformation $\nabla_\mu\Lambda$. This condition ensures that parallel transport develops smoothly as any path is traversed along the base manifold ([C3]).

Notice that Λ in the gauge transformation (4.13) is not constant but an arbitrary function of position and time. Changing the value of the electrostatic potential V by a constant amount is an example of what we have called a global transformation in Section 3.5 (since the change in the potential is the same everywhere). Invariance under this global transformation is related to a conservation law: that of electric charge.

But this global invariance is not sufficient to generate the full Maxwellian dynamics. However, one can regard equations (4.12) and (4.13) as expressing the fact that the local change in the electrostatic potential V (the $\partial\Lambda/\partial t$ term in (4.13)) can be compensated – in the sense of leaving Maxwell's equations unchanged – by a corresponding local change (4.12) in the magnetic vector potential $\boldsymbol{A}$. Thus by including magnetic effects, the global invariance under a change of V by a constant can be extended to a local invariance (4.19) – which is a much more restrictive condition to satisfy ([A1]).

If a 4-vector potential A_μ is postulated, and it is then demanded that the theory involve it only in a way that is insensitive to local changes of the form (4.19), one is led naturally to the idea that the physical fields enter only via the quantity $F_{\mu\nu}$, which is invariant under (4.19).

The idea that dynamics (in this case, the complete interconnection of electric and magnetic effects) may be intimately related to a local invariance requirement (in this case, electromagnetic gauge invariance) turns out to be a fruitful one. It is generally the case that, when a certain global invariance is generalized to a local one, the existence of a new 'compensating' field is entailed, interacting in a specified way. The first example of dynamical theory 'derived' from a local invariance requirement seems to be the theory of Yang and Mills, as we shall see in detail in Chapter 6 ([A1]).

The electromagnetic gauge invariance (4.19) is in the quantum form of the theory directly related to an invariance under *local phase transformations*. It means that the gauge transformation (4.19) corresponds to a unitary operator that transforms the wave function of a particle in an electromagnetic field ([A1]).

A full understanding of gauge invariance in electrodynamics can only be reached via the formalism of quantum field theory, which is not easy to master. Nevertheless, many of the crucial ideas can be discussed within the more familiar framework of ordinary quantum mechanics, rather than quantum field theory, treating electromagnetism as a purely classical field ([A1]) ([100]). Thus in order to proceed further, the next thing that we have to discuss is how such (gauge) ideas are incorporated into (ordinary) quantum mechanics.

But before we continue, let us explain why actually the term 'gauge' is used.

4.5. What is 'gauge' in a gauge theory?

The history of gauge theories begins with general relativity (GR), which can be regarded as a (non-Abelian) gauge theory of a special type. To a large extent, the other gauge theories emerged in a slow and complicated process gradually from GR ([S15]).

It all began with H. Weyl [W5], who made in 1918 the first attempt to extend general relativity (GR) in order to describe gravitation and electromagnetism within a unifying geometrical framework. This brilliant proposal contains the germs of all mathematical aspects of non-Abelian gauge theory. The word 'gauge' (German: 'Eich-') transformation appeared for the first time in this paper but in the everyday meaning of a change of length or change of calibration (like the 'gauge' of railway tracks) ([101]).

---

[100] In this paper I discuss Yang-Mills theory up to the point of its quantisation. It is, of course, quantisation which lends physical significance the theory. But it opens a can of mathematical worms which have no place in this layman's exposition.

[101] The German word *eichen* probably comes from the Latin *aequare*, i.e., equalizing the length to a standard one ([O3]).

In Euclidean geometry, the length and direction of a vector are invariant with respect to a translation. In Riemann's geometry, however, only the length remains unchanged. Weyl wondered why it is so. Ultimately, the measuring devices have to be transported from one point in spacetime to another in order to measure the length (and time) there. According to Weyl, therefore, one can only measure the relative lengths of two vectors. The absolute length of a vector, however, is arbitrary.

To describe this mathematically, Weyl modified the Einstein spacetime metric $g_{\mu\nu}$ as follows:

$$T_\lambda : g_{\mu\nu} \to g'_{\mu\nu} = \lambda g_{\mu\nu},$$

where the 'conformal factor' $\lambda = \lambda(t, \boldsymbol{x})$ is any positive and smooth ([102]) function on spacetime. Weyl found out that general relativity is invariant with regard to this transformation, i.e. the 'gauge' λ can be chosen as an arbitrary function. He called the transformation T_λ 'gauge transformation' (of the length).

In other words, in addition to the requirement of the invariance of the physical laws against arbitrary coordinate transformations (according to general relativity), Weyl also put the invariance with regard to this 'gauge transformation' on an equal footing and described this as the *principle of gauge invariance*.

One can show that there is a 4-vector field ϕ_μ on spacetime, such that the transformation T_λ corresponds to the following ('gauge') transformation:

$$\phi_\mu \to \phi_\mu - 1/\lambda \nabla_\mu \lambda.$$

Since λ is arbitrary, we can set $\lambda = -e^\alpha$, where $\alpha = \alpha(t, \boldsymbol{x})$ is a (smooth) function on spacetime. The transformation $g_{\mu\nu} \to \lambda g_{\mu\nu}$ then looks like this: $\phi_\mu \to \phi_\mu - \nabla_\mu \alpha$, i.e. it looks exactly like the transformation (4.19) of the electromagnetic potential field $A_\mu \to A_\mu - \nabla_\mu \Lambda$.

Weyl then postulated that there is only one vector field ϕ_μ having physical meaning, namely $\phi_\mu = e/\gamma A_\mu$, where e is the elementary charge (and γ an indefinite constant). This enabled Weyl to unify gravitation and electromagnetism, assuming that the 'gauge' can be arbitrarily chosen locally at points of spacetime.

In this way, Weyl replaced Riemann's geometry with another one, which is now called Weyl's geometry. If in this geometry, a vector with length L_P is transported from a point P in spacetime to a point Q in parallel, its length L_Q in Q is dependent on the path K and the (electromagnetic) potential field A_μ:

$$L_Q = L_P \exp(e/\gamma \int_K A_\mu).$$

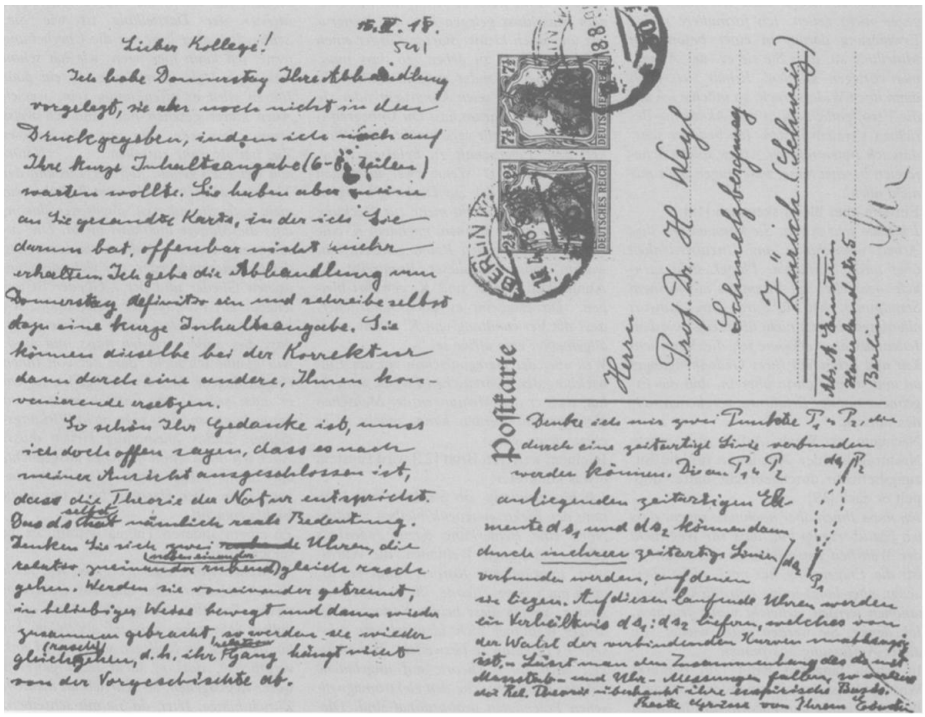
The further history of the Weyl theory is well known: Einstein immediately raised an objection to it (although the mathematics of the theory impressed him very much). He argued that according to the 'Weyl's gauge invariance', the size (or the rate speed) of a watch would change when the watch moves through an electromagnetic field. This would mean that two synchronized clocks would no longer be synchronized after moving in the electromagnetic field, which contradicts known observations.

Einstein: "... *As nice as your thought is, I have to say frankly that in my opinion it is impossible that the theory corresponds to nature*". ([103]) ([P9])

---

[102] A smooth function is a function that has continuous derivatives up to some desired order. The number of continuous derivatives necessary for a function to be considered smooth depends on the problem at hand, and may vary from two to infinity.

[103] „... *So schön Ihr Gedanke ist, muss ich doch offen sagen, dass es nach meiner Ansicht ausgeschlossen ist, dass die Theorie der Natur entspricht*".



This led to an intense exchange of letters between Einstein (in Berlin) and Weyl (at the ETH in Zürich). No agreement was reached, but Einstein's intuition proved to be right. After a long discussion, Weyl finally admitted that his attempt was a failure as a physical theory (104) ([O3], [S13]).

Although Weyl's attempt was a failure as a physical theory, it paved the way for the correct understanding of gauge invariance which plays a very important role in modern physics. But that had to wait for quantum mechanics to come. Weyl himself reinterpreted in 1929 his original theory after the advent of quantum theory in a seminal paper [W7], this time relating electromagnetism not to gravity but to the wave-field of the electron (105). The role of the metric is taken over by the wave function, and the rescaling of the metric was replaced by a phase change of the wave function. Weyl's reinterpretation of his earlier speculative proposal had actually been suggested before by F. London (1927), V. Fock (1926), O. Klein (1926), and others ([S15]).

Today, so-called gauge theories have nothing to do with geometrical objects like $g_{\mu\nu}$; instead, they involve local phase changes in quantum fields, i.e. $\psi \rightarrow e^{i\Lambda(t,\boldsymbol{x})}\,\psi$ (106), which are

104 In his encyclopaedia article on relativity [P2] Wolfgang Pauli commented on Weyl's point of view ([S14]): "... *In summary one may say that Weyl's theory has not yet contributed to getting closer to the solution of the problem of matter*."

105 Before its release, Weyl published a short summary to which Pauli, upset by the mathematician's intrusion into physics, replied: "... *I admire your courage; since the conclusion is inevitable that you wish to be judged, not for your success in pure mathematics, but for your true but unhappy love for physics*." However, after reading the whole article [W7] Pauli became more friendly and wrote ([S14]): "... *Here I must admit your ability in Physics. Your earlier theory with* $g_{\mu\nu}$ *was pure mathematics and unphysical. Einstein was justified in criticizing and scolding. Now the hour of your revenge has arrived*."

106 Notice that $e^{i\Lambda(t,\boldsymbol{x})}$ is a complex number. Recall that a *complex number* is a number of the form $z = x + iy$, where x and y are real numbers, and i is a symbol called the *imaginary unit*, and satisfying the equation $i^2 = -1$. The absolute value of z (or *modulus* or *magnitude*) is $r = |z| := \sqrt{x^2 + y^2}$. The magnitude of a complex number $z = x + iy$ is obviously the Euclidean distance from the origin in the complex plane to the point (x, y). The complex *conjugate* of the complex number $z = x + iy$ is given by $x - iy$. It is denoted by either $\bar{z}$ or z*. Clearly $zz^* = x^2 + y^2 = |z|^2 = |z^*|^2$. The angle φ or *phase* (or *argument*) of the complex number $x + iy$ is the angle, measured in radians, from the point $1 + i0$ to $x + iy$, with counter clockwise denoting positive angle.
We can express the complex number $z = x + iy$ in terms of the *complex exponential* as $z = re^{i\varphi}$, where $r = |z|$. It is called the *exponential* or *polar* of the complex number z (since (r, θ) are the polar coordinates of the point with rectangular coordinates (x, y)). Recovering the original rectangular coordinates from the polar form is done by the formula called *trigonometric form* $z = r(\cos\varphi + i\sin\varphi)$. The set of all complex numbers is denoted by $\mathbb{C}$ ([W11]).

fundamental in the description of the electroweak and strong interactions in the Standard Model. The notion of *gauge symmetry* is used for historic reasons and does not make much sense for the type of symmetry we are considering here.

5. Gauge invariance in quantum mechanics (QM)

In the present chapter ([107]) we shall discuss gauge invariance in electrodynamics within the framework of ordinary quantum mechanics, rather than quantum field theory, treating electromagnetism as a purely classical field. We will see that in the quantum form of the theory the gauge invariance of Maxwell's equations is directly related to invariance under local phase transformations of the (complex-valued) wave function ψ of a particle with the charge q.

We begin with non-relativistic quantum mechanics and then we shall explore the generalization to relativistic quantum mechanics, for particles of spin-0 (via the Klein–Gordon equation) and spin-½ (via the Dirac equation) ([A1]).

5.1. An interlude on Lagrangian formalism

The majority of contemporary field theories, including the quantum fields of the Standard Model of particle physics, are derived from an action principle, a fundamental concept in theoretical physics. In essence, the action principle posits that a system is defined by an expression, termed the *Lagrangian*, an equation that captures the presumed kinetic and potential energies of the system in a specific form. The system's evolution from an initial to a final state is such that the integral of this Lagrangian, known as the *action*, is minimised. From the Lagrangian, the principle of least action allows the derivation of other equations, including the equations of motion or field equations of the system.

This formalism is a formulation of classical mechanics, wherein the dynamics of the system under consideration is described by a single scalar function, the Lagrangian density. All field theories can be described with mathematical formulas of the Lagrangian densities. The Lagrangian density is thus the fundamental quantity of a field theory.

The Lagrangian formulation of classical field (or particle) physics is a powerful approach due to its ability to efficiently incorporate symmetries and demonstrate their connection with conservation laws ([108]). This is also the case in the context of quantum field theory (QFT). A QFT is defined in terms of its Lagrangian, which in turn defines the fields and their interactions.

Thus for the understanding of a field theory (like e.g. electrodynamics) the Lagrange ([109]) formalism is of central importance. It provides a unified framework for the mechanics of both fields and particles, encompassing both classical and quantum aspects. It is noteworthy that this framework was developed prior to the advent of quantum field theory. This was in the form of the *principle of least action* ([110]), with the action defined in terms of a Lagrangian ([A1]) ([111]).

The *Lagrangian density* $\mathcal{L}$ is a (real) function of all fields ϕ_i of the theory and their derivatives $\nabla_\mu \phi_i$:

---

[107] Many fragments of this and the next two chapters are borrowed from the book [A1] that I highly recommend to the reader for further reading. [A1] and [S2] are the clearest introductory books I have found.

[108] Any symmetry (both internal and external) of the Lagrangian leads to a conservation law: e.g., spatial translation symmetry leads to momentum conservation, whereas symmetry under spatial rotations leads to the conservation of angular momentum.

[109] Joseph-Louis de Lagrange (1736 - 1913), also reported as Giuseppe Luigi Lagrange or Lagrangia was an Italian mathematician and astronomer, later naturalised French.

[110] The principle of least action had been proposed in various forms by Pierre Fermat (1601–1665) and Pierre-Louis Moreau de Maupertuis (1698–1759). It was later developed and formalized by Leonhard Euler (1707–1783) and Lagrange.

[111] The reader is referred to a great exposition of the topic in [F4].

$$\mathcal{L} = \mathcal{L}(\mathrm{t}, \phi_i, \nabla_\mu \phi_i).$$

The explicit form of the Lagrangian density depends on the field theory under consideration. The spatial volume integral of the Lagrangian density $\mathcal{L}$ is called the *Lagrangian* ([112])

$$L := \int \mathcal{L}\, d^3x$$

The time integral of the Lagrangian is called the *action* denoted by S

$$S := \int L\, dt = \int \mathcal{L}(\mathrm{t}, \phi_i, \nabla_\mu \phi_i)\, d^4x_\mu.$$

The action is often referred to as the *action functional*, in that it is a function of the fields (and their derivatives).

Now we postulate the fundamental Hamilton's *principle of least action*

(5.1) $$\delta S = 0,$$

where δS is the *variation* of S, meaning that the dynamics of the fields is governed by minimizing ([113]) the action S. This principle leads to the *Euler-Lagrange* equations of field theory. These equations form a system of ordinary differential equations (second order with respect to the time derivative) that uniquely determine the behaviour of the fields. The number of differential equations depends on the number of degrees of freedom.

Let us explain it in more detail. As a rule, the Lagrangian is a function of the position and velocity along the trajectory of a system. In simple situations (in classical mechanics) the Lagrangian is just the difference $L = T - V$, where $T =$ kinetic energy of the system and $V =$ potential energy of the system. Yes, there is a minus sign in the definition (a plus sign would simply give the total energy, i.e. the Hamiltonian). The kinetic energy T is usually a function of the velocities, while the potential energy V is a function of the positions.

Suppose we have a single (classical) particle with mass m (in a gravitational field, for instance) which starts somewhere and moves to some other point by free motion. In the Newtonian approach, the trajectories of the particle are calculated using equations of motion which involve forces as the physical input. In the least action approach, the path by which a particle actually travels is determined by the principle that it has to follow that particular path, out of infinitely many possible ones, for which the variation of the action S vanishes (i.e. $\delta S = 0$) ([114]).

For a single particle in a potential V the Lagrangian L is given by

$$L = L(x(t), \dot{x}(t)) = \frac{1}{2} m\dot{x}^2 - V(x),$$

where $x(t)$ is the position of the particle as a function of time, $\dot{x}(t)$ is its velocity, i.e. $\dot{x}(t) = \frac{dx}{dt}$. Thus the action S is given by

$$S = S(x(t)) = \int_{t_1}^{t_2} L dt = \int_{t_1}^{t_2} \left[\frac{m}{2}\left(\frac{dx}{dt}\right)^2 - V(x)\right] dt.$$

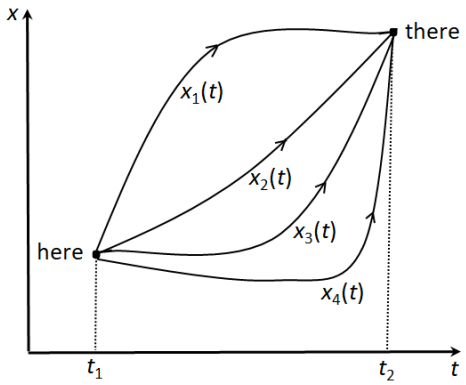
Since S is an explicit function of the variables x and $\dot{x}$, so knowing $V(\mathrm{x})$ we can evaluate S for all sorts of possible $x(t)$'s starting at time t_1 and ending at time t_2. We can draw these different possible trajectories on a x versus t diagram as in the following figure

---

[112] It is a customary abuse of language in quantum theory to refer to the 'Lagrangian density' as the 'Lagrangian'.

[113] We say *least* or *minimum*, but what is really meant is *stationary*, i.e. maximum or minimum. $\delta S = 0$ basically means that a slight variation (differential) in the action should be zero. For this reason purists prefer the name *principle of stationary action*. ([H13])

[114] Now, the reader may wonder why exactly should physical systems obey this principle. Why should an object take the path of stationary action instead of some other path? Fundamentally, nobody actually knows the real answer to this.



For each path we evaluate S: the actual path is the one for which S is smallest, by the principle of least action (5.1). Of course, the path $x_0(t)$ that has the minimum action is in this case the one satisfying Newton's second law. It can be obtained analytically by solving the *Euler-Lagrange* equation of motion

$$\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = 0.$$

Indeed, the Euler-Lagrange equation amounts in our case to

$$m\ddot{x} = -\frac{dV(x)}{dx},$$

where $\ddot{x} = \frac{d^2x}{dt^2}$. But $-dV/dx$ is the force F on the particle. So we see that the Euler-Lagrange equation says exactly the same thing as Newton's second law: $F = ma$, where $a = \ddot{x}$.

In a field theory with a Lagrangian density $\mathcal{L}(t, \phi_i, \nabla_\mu \phi_i)$ the Euler-Lagrange equation has the form ([115])

$$\frac{\partial \mathcal{L}}{\partial \phi_i} - \nabla_\mu\left(\frac{\partial \mathcal{L}}{\partial(\nabla_\mu \phi_i)}\right) = 0.$$

The Euler-Lagrange equation can be used for any given Lagrangian density $\mathcal{L}$ to derive the corresponding equation of motion ([116]). Solutions of this equation of motion correctly describe how a system evolves.

As we already know, a relativistic field theory must be covariant under the Lorentz transformations. Therefore, the action (i.e. also the Lagrangian density) must also satisfy the Lorentz invariance. The Lagrangian density (or at least the action) of a gauge theory should additionally be invariant under the gauge transformation.

The Lagrangian density, called *kinetic term*, for the free (i.e. non-interacting) electromagnetic field is (cf. (4.9))

$$\mathcal{L}_{em}^{free} := -\frac{1}{4\mu_0} F_{\mu\nu}F^{\mu\nu} = -\frac{1}{2\mu_0}(\boldsymbol{B}^2 - \boldsymbol{E}^2),$$

where μ_0 is the permeability of free space, a physical constant connected to the energy stored in a magnetic field ([117]). The interaction between the charged particle (or in general any charged body) with some charge density j_μ and an electromagnetic field is given by the Lagrangian density called *interaction term*

---

[115] More precisely, we get one equation for each field component ϕ_i.

[116] The reader may ask where to get a Lagrangian density from. In general, the answer is the same as in classical physics: derive it from experiment, guess, borrow it from some theory that you like, or just pick one and see what it does. ([S17])

[117] It is related to the speed of light by the equation $c = 1/\sqrt{\mu_0 \varepsilon_0}$. Recall that ε_0 is the electric constant.

$$\mathcal{L}_{em}^{int} := -\mathrm{j}^{\mu} A_{\mu}. \ (118)$$

This term describes interaction potential energy. All together, the Lagrangian density of a charged particle and an electromagnetic field is (119):

(5.2) $$\mathcal{L}_{em} := \mathcal{L}_{em}^{free} + \mathcal{L}_{em}^{int} = -\frac{1}{4\mu_0} F_{\mu\nu} F^{\mu\nu} - \mathrm{j}^{\mu} A_{\mu}.$$

It is immediate that $\mathcal{L}_{em}$ is Lorentz invariant (see Section 4.3). But is it gauge invariant? At first sight, this Lagrangian density is not gauge invariant, because $A'_{\mu} = A_{\mu} - \nabla_{\mu}\Lambda$ implies

$$\mathrm{j}^{\mu} A'_{\mu} = \mathrm{j}^{\mu} A_{\mu} - \mathrm{j}^{\mu} \nabla_{\mu} \Lambda \neq \mathrm{j}^{\mu} A_{\mu}.$$

However the extra term $\mathrm{j}^{\mu}\nabla_{\mu}\Lambda$ can be transformed using the continuity equation (4.2) into the divergence $\nabla_{\mu}\mathrm{j}^{\mu}\Lambda$ which can be discarded if the function Λ fulfils reasonable boundary conditions (120). The Euler-Lagrange equations corresponding to the Lagrangian density $\mathcal{L}_{em}$ become Maxwell's field equations.

We close this section with the following remark. In the example above, we considered (infinitely) many possible paths $x_i(t)$, of which only one was the actual path followed by a (classical) particle, namely the one we called $x_0(t)$ which minimized the action integral as a functional of $x(t)$. In the quantum case, however, a particle will no longer follow any definite path, because of quantum fluctuations. In the original formulations of quantum theory, such fluctuations were generally taken to imply that the very notion of a 'path' was no longer a useful one. Fortunately, a quantum generalization of the 'path-contribution to the action' approach to classical mechanics was subsequently developed. The idea was first hinted at in 1933 by Dirac, but it was Feynman who worked it out completely (121). ([A1])

Instead of throwing away the whole idea of a path, Feynman's insight was to consider the 'opposite' viewpoint: since unique paths are forbidden in quantum theory, we should in principle include *all possible* paths. However, surely not all paths are equally likely: after all, we must recover the classical trajectory as a limit. Thus we must find an appropriate weighting for the paths. Feynman's recipe is beautifully simple: the contribution of a path is proportional to $e^{\mathrm{i}S/\hbar}$, where S is the action for that particular path. In summary, the quantum mechanical amplitude to go from $x(t_1)$ to $x(t_2)$ is proportional to

$$\sum_{all\ paths\ x(t)} exp\left(\mathrm{i}\int_{t_1}^{t_2} L\big(x(t), \dot{x}(t)\big)dt\right).$$

---

118 Reminder: $\mathrm{j}^{\mu} := \eta^{\nu\mu}\mathrm{j}_{\mu} = (\rho, -\boldsymbol{j})$, where $\eta^{\nu\mu}$ is the Minkowski metric and j_{μ} is the electromagnetic 4-current density.

119 We ignore kinetic energy $-\frac{1}{2}mv_{\mu}v^{\mu}$ of the particle with mass m in this Lagrangian density.

120 It means that we restrict to gauge transformation or to currents that vanish sufficiently fast at infinity.

121 Dirac's work did not provide a precise prescription to calculate the sum over paths, and he did not show that one could recover the Schrödinger equation or the canonical commutation relations from this rule. This was done by Feynman. That is, the classical path arises naturally in the classical limit.
In his paper, Dirac had cryptically remarked that a critical quantum quantity is 'analogous' to its classical counterpart, but Feynman believed that the correct phrase was 'proportional to'. In September 1946, at a conference in Princeton, Feynman had an opportunity to find out what Dirac had meant. He described his problem to Dirac and came to the crunch:

Feynman: *Did you know that they were proportional?*
Dirac: *Are they?*
Feynman: *Yes they are.*
Dirac: *That's interesting.*
After a silence Dirac walked away. ([F1])

Niels Bohr described Dirac as "*the strangest man*". His extreme reticence, monosyllabic responses and repetitious statements are legendary. It has been said in jest that Dirac's spoken vocabulary consisted of *"Yes"*, *"No"*, and *"I don't know"*.

We shall not, however, make use of the 'path integral' approach to quantum field theory in this paper. Its use was, in fact, crucial in obtaining the Feynman rules for non-Abelian gauge theories; and it is the only approach suitable for numerical studies of quantum field theories ([A1]).

5.2. The Schrödinger equation

Maxwell's equations represent a classical and not a quantum mechanical description of electromagnetic waves. In contrast, the Schrödinger equation in quantum mechanics is a wave equation for quantum mechanical particles.

The Schrödinger equation was formulated in 1926 by the Austrian physicist Erwin Schrödinger (1887–1961) as a wave equation and is the fundamental equation of quantum mechanics ([122]). In the form of a partial differential equation, it describes the development of a quantum state over time in a non-relativistic quantum mechanical system, say a bunch of particles subject to certain forces ([123]).

In this framework, the scalar (complex-valued) field $(\mathrm{t}, \boldsymbol{x}) \to \psi(\mathrm{t}, \boldsymbol{x}) \in \mathbb{C}$ ([124]) of a non-relativistic spin-0 particle with mass m under the influence of a potential $V(\mathrm{t}, \boldsymbol{x})$ is determined by the (time-dependent) Schrödinger equation (reminder: $\hbar = 1$):

$$\mathrm{i}\frac{\partial}{\partial t}\psi = [-\frac{1}{2m}\nabla^2 + V(\mathrm{t}, \boldsymbol{x})]\psi. \tag{5.3}$$

The starting point for Schrödinger was the classical Hamiltonian (which corresponds to the total energy $E = T + V$) for a particle of mass m, which moves with momentum $\boldsymbol{p}$ in a potential V ([125])

$$E = \frac{\boldsymbol{p}^2}{2m} + V(\mathrm{t}, \boldsymbol{x}).$$

Multiplying both sides by ψ we get

$$E\psi = [\frac{\boldsymbol{p}^2}{2m} + V(\mathrm{t}, \boldsymbol{x})]\psi.$$

Schrödinger replaced then the (classical) quantities energy E, momentum $\boldsymbol{p}$ and position $\boldsymbol{x}$ with the following quantum operators:

$$E \to \mathrm{i}\hbar\partial/\partial t, \quad \boldsymbol{p} \to -\mathrm{i}\hbar\nabla, \quad \boldsymbol{x} \to \boldsymbol{x} \ ([126])$$

and obtained

---

[122] *"Schrödinger's revolutionary work on the wave mechanics version of quantum mechanics came together in the last weeks of December 1925. He was staying in an inn up in the mountains at Arosa with a girlfriend from Vienna, one whose identity remains a mystery to this day. He was working on what was to become known as the 'Schrödinger equation'.* (…) *When he got back to Zürich, he consulted [his closest friend] Weyl, who was an expert on this kind of equation, and explained to him what the general properties of its solutions were. In his first paper on quantum mechanics, Schrödinger explicitly thanks Weyl for his help. (…)*
Weyl later commented on this period by remarking that Schrödinger 'did his great work during a late erotic outburst in his life'. Schrödinger was married, but was 'convinced that Bourgeois marriage, while essential for a comfortable life, is incompatible with romantic love'. His wife, Annemarie, presumably was not too concerned about his spending time in the mountains with his girlfriend, since she was Weyl's lover at the time." ([W13]).

[123] The equation has now been confirmed by countless experiments. However, it is only applicable for certain situations, e.g. if one can ignore the spin of vector particles.

[124] The position-space wave function ψ for N particles is of the form $\psi(\mathrm{t}, \boldsymbol{x}_1, ..., \boldsymbol{x}_N)$, where $\boldsymbol{x}_i$ is the position of the i^{th} particle in three-dimensional space, and t is time. Altogether, this is a complex-valued function of 3N+1 real variables, i.e. ψ is defined on the 3N+1 dimensional configuration space.

[125] Physically, the quantity $\frac{\boldsymbol{p}^2}{2m}$ represents kinetic energy, while $V(\mathrm{t}, \boldsymbol{x})$ represents potential energy. Notice that given the Lagrangian for a system, one can construct the Hamiltonian and vice versa: the Lagrangian is just the Legendre transform of the Hamiltonian.

[126] The reader may well wonder why *this* replacement is used. We shall postpone the discussion of this question to Section 5.9.

$$\mathrm{i}\hbar\frac{\partial}{\partial t}\psi = [\frac{(-\mathrm{i}\hbar\nabla)^2}{2m} + V(\mathrm{t}, \boldsymbol{x})]\psi,$$

which is exactly equation (5.3) since we assume $\hbar = 1$. These operators act now on an unknown wave function ψ. This procedure is called the *first quantisation* (or *canonical quantisation*) ([127]). Note that for historical reasons, the energy operator on the right-hand side of the Schrödinger equation is known as the *Hamiltonian operator* (cf. Section 5.9)

$$\widehat{H} := [\frac{(-\mathrm{i}\hbar\nabla)^2}{2m} + V(\mathrm{t}, \boldsymbol{x})].$$

Therefore, in terms of the Hamiltonian operator, the Schrödinger equation reads:

(5.3') $$\mathrm{i}\hbar\frac{\partial}{\partial t}\psi = \widehat{H}\psi.$$

After postulating equation (5.3) there was a great deal of discussion on what the wave function actually meant. A solution to the Schrödinger equation is the (quantum) wave function $\psi = \psi(\mathrm{t}, \boldsymbol{x})$. But how is this solution to be interpreted physically? Schrödinger himself gave the first interpretation based on classical physics. The wave optics links the intensity of a wave with the square of the amplitudes. The question therefore arose whether the quantity $|\psi|^2 = \psi\psi^*$ ([128]) could also be assigned to a measurable property of the quantum object whose state ψ described. However, this interpretation of the wave function led to contradictions with experiments that showed that the wave function cannot be understood as a field function in the classical sense.

After much debate, the wave function is now accepted to be a probability distribution (density) ([129]). According to this statistical interpretation of quantum mechanics proposed by Max Born ([130]) in 1926, the (squared) absolute value $|\psi(\mathrm{t}, \boldsymbol{x})|^2$ corresponds to the probability that the particle will be found in the position $\mathrm{x} = (\mathrm{t}, \boldsymbol{x})$ ([131]). To do this, however, the wave function must be normalised in such a way that the total probability is one:

$$\int_{\mathbb{R}^3} |\psi(t, \boldsymbol{x})|^2 \, d^3x = 1.$$

In particular, the probability integral must be time independent. This statistical interpretation takes into account the fact that the experimental results are mean values over many individual events.

The Schrodinger equation is used to find the allowed energy levels of quantum mechanical systems (such as atoms). The associated wave function gives the probability of finding the particle at a certain position.

In general, the wave function does not describe a physical wave because it is not a function defined on physical space. Rather, it is defined on *configuration space* – it takes as input all the

---

[127] The term first quantisation refers to its relationship to the second quantisation (= field quantisation). Historically, it was not the first attempt of quantisation in modern physics.

[128] The complex conjugate ψ^* of ψ is defined by $(\mathrm{t}, \boldsymbol{x}) \to \psi^*(\mathrm{t}, \boldsymbol{x}) := (\psi(\mathrm{t}, \boldsymbol{x}))^*$. Recall that every complex number $\mathrm{z} = \mathrm{x} + \mathrm{iy}$ (where $\mathrm{i}^2 = -1$) has associated with it the *complex conjugate* $\mathrm{z}^* := \mathrm{x} - \mathrm{iy}$. Thus the product $\mathrm{zz}^* = \mathrm{z}^*\mathrm{z} = \mathrm{x}^2 + \mathrm{y}^2$ is equal to the squared magnitude of z: $|\mathrm{z}| = \sqrt{x^2 + y^2}$.

[129] More accurately, wave function is not probability itself; it is probability amplitude. This means, you have to take its absolute square in order to get a probability. So just remember: probabilities are real numbers between 0 and 1 (or between 0 and 100%), while amplitudes are complex numbers.

[130] Max Born (1882-1970) was a German physicist and mathematician. He won the 1954 Nobel Prize in Physics.

[131] To be more precise, in the context of quantum field theory (QFT), the concept of particles is rendered obsolete, as the theory posits the existence of fields alone. Within the framework of QFT, interactions, including creation and destruction, occur at specific locations x. However, the fundamental objects of the theory, namely the fields, are said to lack positions due to their infinitely extended nature. Rigorous analysis demonstrates that, even under the broadest definition of "particle," particles are incompatible with the combined principles of relativity and quantum physics. This analysis further demonstrates that photons, in particular, cannot be considered point particles; relativistic and quantum principles imply that a photon cannot be "located" at a specific point, even in principle. ([H7])

possible configurations of locations the particles could be in and it returns a value related to the probability that you will find the particles in the given configuration at the given time ([132]).

Another aspect of the Schrödinger equation is its linearity. Unlike the classical equations, which are nonlinear, the Schrödinger equation is linear. This linearity gives quantum mechanics some of its uniquely non-classical characteristics, such as the superposition of states ([Z6]). This means that if some wave function ψ_1 is a solution and some other wave function ψ_2 is also a solution, then the sum $\psi = \psi_1 + \psi_2$ is a solution too. But ψ_1 and ψ_2 could correspond to quite different situations, for example ψ_1 might correspond to the particle being in a laboratory A and ψ_2 might correspond to the particle being in a distant laboratory B. Since the sum $\psi_1 + \psi_2$ is also a solution, there is a sense in which the **same** particle is in both places at once. When this happens we say that the particle is in *superposition* of the two states ψ_1 and ψ_2. ([133])

We now look more closely into the relationship between the gauge invariance of Maxwell's equations and invariance under local phase transformations in quantum mechanics.

First let us notice that the free-particle Schrödinger equation (i.e. for $V(\mathrm{t}, \boldsymbol{x}) = 0$) amounts to

$$\mathrm{i}\partial\psi/\partial t = -\frac{1}{2m}\nabla^2\psi. \tag{5.4}$$

The Schrödinger equation for a spin-0 particle of charge q (e.g. charged pion π^+) in an electromagnetic field is

$$\mathrm{i}\partial\psi/\partial t = [\frac{1}{2m}(-\mathrm{i}\boldsymbol{\nabla} - q\boldsymbol{A})^2 + qV]\psi, \tag{5.5}$$

where $\boldsymbol{A}$ is the magnetic 3-potential and V the electric scalar potential (Section 4.4). We can write (5.5) as

$$\mathrm{i}(\partial/\partial t + \mathrm{i}qV)\psi = -\frac{1}{2m}(\boldsymbol{\nabla} - \mathrm{i}q\boldsymbol{A})^2\psi. \tag{5.6}$$

Note the appearance of the operator combinations

$$D_t := \partial/\partial t + \mathrm{i}qV \quad \text{and} \quad \boldsymbol{D} := \boldsymbol{\nabla} - \mathrm{i}q\boldsymbol{A} \tag{5.7}$$

in place of $\partial/\partial t$ and $\boldsymbol{\nabla}$, in going from the free-particle Schrödinger equation (5.4) to the electromagnetic field case (5.6). We note that (5.7) can be written in manifestly Lorentz invariant form as ([134])

$$D_\mu := (D_t, -\boldsymbol{D}) = \nabla_\mu + \mathrm{i}qA_\mu. \tag{5.8}$$

---

[132] While the wave function generally does not represent a straightforward wave in three-dimensional space, the question remains whether there is some sort of physical wave associated to it. Several physicists, including Einstein, de Broglie, Schrödinger and Bohm, believed that there should be, but although efforts to find one still continue today, they have not resulted in theories that enjoy mainstream approval.
Others, including Pauli, Heisenberg and Bohr were against this realistic picture and regarded the wave function as a mere mathematical tool to provide probabilities. Indeed, they argued that questions such as "where is the particle when we are not looking" are meaningless: science cannot describe nature per se, but only our knowledge of it. So the only kind of questions we can answer are questions about possible outcomes of measurements. And that is precisely what the wave function gives us. This view is known as the *Copenhagen interpretation* of quantum mechanics. It is in stark contrast to the intuition classical physics is based on: that there exists an objective reality even when we are not looking and that science can describe that reality ([F16]).
For an in-depth discussion of this issue the reader is referred to [S9].

[133] It is because of this interference that we have to add probability amplitudes first, before we can calculate the probability of an event happening in one or the other (indistinguishable) way (let us say A or B) – instead of just adding probabilities as we would do in the classical world. It makes a big difference: $|\psi_A + \psi_B|^2$ is the probability when we cannot distinguish the alternatives, while $|\psi_A|^2 + |\psi_B|^2$ is the probability when we can see what happens (i.e. we can see whether A or B was the case). Now, $|\psi_A + \psi_B|^2$ is definitely not the same as $|\psi_A|^2 + |\psi_B|^2$. ([R2])

[134] Reminder: $\nabla_\mu = (\partial/\partial t, -\boldsymbol{\nabla})$ and $A_\mu = (V, \boldsymbol{A})$. D_μ is manifestly invariant because all the objects in its definition are 4-vectors.

The solution $\psi(t, \boldsymbol{x})$ of the Schrödinger equation (5.5) describes completely the state of the particle moving under the influence of the potentials V and $\boldsymbol{A}$. However, these potentials are not unique, as we have already seen: they can be changed by a gauge transformation (cf. (4.12) and (4.13))

$$\boldsymbol{A} \to \boldsymbol{A}' = \boldsymbol{A} + \nabla\Lambda, \qquad V \to V' = V - \partial\Lambda/\partial t \tag{5.9}$$

without changing Maxwell's equations for the fields $\boldsymbol{E}$ and $\boldsymbol{B}$.

This immediately raises an important question: if we carry out such a change of potentials in equation (5.5), will the solution of the resulting equation describe the same physics as the solution of equation (5.5)? If it does, we shall be able to assume the validity of Maxwell's theory for the quantum world; if not, some modification will be necessary, since the gauge symmetry possessed by Maxwell's equations will be violated in the quantum theory. ([A1])

The answer to this question is obviously negative since the same ψ cannot possibly satisfy both (5.5) and the analogous equation with $(V, \boldsymbol{A})$ replaced by $(V', \boldsymbol{A}')$. Unlike Maxwell's equations, the Schrödinger equation is not gauge invariant. But we must remember that the wave function ψ is not a directly observable quantity, as the electromagnetic fields $\boldsymbol{E}$ and $\boldsymbol{B}$ are. Perhaps ψ does not need to remain unchanged (invariant) when the potentials are changed by a gauge transformation. In fact, in order to have any chance of 'describing the same physics' in terms of the gauge-transformed potentials, we will have to allow ψ to change as well. This is a crucial point: for quantum mechanics to be consistent with Maxwell's equations it is necessary for gauge transformations (5.9) of the electromagnetic potentials to be accompanied also by a transformation of the quantum-mechanical wave function $\psi \to \psi'$. ([A1])

It turns out that the required ψ' is

$$\psi'(t, \boldsymbol{x}) := e^{iq\Lambda(t,\boldsymbol{x})}\psi(t, \boldsymbol{x}), \tag{5.10}$$

where Λ is the same spacetime-dependent function as appears in equations (5.9). Indeed, it is easy to verify that when we replace in (5.5) $(V, \boldsymbol{A})$ and ψ by $(V', \boldsymbol{A}')$ and ψ', respectively, then the form of the resulting equation is exactly the same as the form of (5.5). Thus it means that (5.5) is *gauge covariant*, i.e. it maintains the same form under the combined transformation

$$\boldsymbol{A} \to \boldsymbol{A}' = \boldsymbol{A} + \nabla\boldsymbol{\Lambda}, \;\; V \to V' = V - \partial\Lambda/\partial t, \quad \psi \to \psi' = e^{iq\Lambda}\,\psi. \tag{5.11}$$

It is important to note that gauge transformation in quantum theory involves both the 4-vector potential $A_\mu = (V, \boldsymbol{A})$ and the particle's wave function ψ.

But do ψ and ψ' describe the same physics, in fact? The answer is yes, but it is not quite obvious. Of course, the probability densities are equal

$$|\psi'|^2 = |e^{iq\Lambda}\psi|^2 = |e^{iq\Lambda}|^2\,|\psi|^2 = |\psi|^2,$$

because

$$|e^{i\alpha}|^2 = e^{i\alpha}e^{-i\alpha} = e^0 = 1$$

for every α. However, we can be interested in other observables, for example, involving the derivative operators $\partial/\partial t$ or $\boldsymbol{\nabla}$. Such quantities need not be invariant under (5.10) because the phase Λ is $(t, \boldsymbol{x})$-dependent ([135]).

The solution is to replace $\partial/\partial t$ and $\boldsymbol{\nabla}$ by D_t and $\boldsymbol{D}$, respectively, according to (5.7). It turns out that any equation involving the operator ∇_μ can be made gauge covariant under the combined transformation (cf. (4.19))

$$A_\mu \to A'_\mu = A_\mu - \nabla_\mu\Lambda, \quad \psi \to \psi' = e^{iq\Lambda}\psi, \tag{5.11'}$$

---

[135] It means that derivatives of ψ' yield terms with derivatives of Λ and one cannot get rid of them. This spoils gauge invariance (or covariance).

if ∇_μ is replaced by $D_\mu = \nabla_\mu + \mathrm{i}qA_\mu$ (see (5.8)) ([136]). In fact, this is a simple prescription for obtaining the wave equation for a particle in the presence of an electromagnetic field from the corresponding free particle wave equation: make the replacement $\nabla_\mu \to D_\mu$. This is the basis of the so-called *gauge principle*.

This new kind of derivative $D_\mu = \nabla_\mu + \mathrm{i}qA_\mu$ turns out to be of fundamental importance, as we shall see below. It is called the *gauge covariant derivative*, the term being usually shortened to *covariant derivative*, due to its covariance under gauge transformations ([A1]). One can easily show that the *commutator* $[D_\mu, D_\nu]$ of two covariant derivatives is

$$[D_\mu, D_\nu] := D_\mu D_\nu - D_\nu D_\mu = \mathrm{i}qF_{\mu\nu}. \tag{5.12}$$

The important property of the covariant derivative is that under combined transformations (5.11'),

$$D_\mu\psi \to \mathrm{e}^{\mathrm{i}q\Lambda(\mathrm{t},\boldsymbol{x})} D_\mu\psi(\mathrm{t}, \boldsymbol{x}),$$

i.e. $D_\mu\psi$ transforms in the same way as ψ in (5.10), even when Λ is a function of x_μ. This property ensures the gauge covariance of wave equations.

Observe that the transformation law (5.9) is a property of the gauge field A_μ, and does not know anything about q, which instead is a parameter related to the field ψ. This particular real number quantifies the electric charge of the entity described by ψ. In fact, this 'entity' will generally be some charged quantum particle, such as an electron or proton, and ψ would then be its quantum-mechanical wave function. The potential A_μ provides the mathematical key to the procedure whereby the electromagnetic field interacts with the matter field ψ. The coupling between A_μ and ψ is obtained using the covariant derivative, which depends on the electric charge q of ψ: $D_\mu\psi = (\nabla_\mu + \mathrm{i}qA_\mu)\psi$. This method of coupling matter to the electromagnetic field is known as the *minimal coupling* ([137]) ([M1]).

The vector potential A_μ was optional in classical electromagnetism but is mandatory in quantum mechanics, because it appears explicitly in the equation of motion. However, it is not directly observable, being determined only up to a gauge transformation. David Bohm ([138]) found this situation curious. He theorised that the strange role of the vector potential was a fatal flaw in the theory. In 1959, he and his student Y. Aharonov ([139]) proposed an experiment to test this hypothesis ([A0]). Reliable experimental verification of the Aharonov-Bohm effect was not achieved until more than twenty years later, at which point the result was precisely as predicted by quantum mechanics ([O4]). This outcome serves to reinforce the fundamental role of the vector potential. It is noteworthy, however, that the vector potential itself remains unobserved, as the experiment does not directly measure $\boldsymbol{A}$ but rather the integral $\oint \boldsymbol{A} \cdot dx$ ([H7]).

Our final remark: the Schrödinger equation is obviously non-relativistic ([140]), but Maxwell's equations are fully relativistic. One might therefore suspect that the prescriptions presented here are actually relativistic as well, and this is indeed the case. Besides, the quarks and leptons ([141]) have spin-½, a degree of freedom absent from the Schrödinger (scalar) wave function. In

---

[136] This can be derived from (4.18) by substituting E with $\mathrm{i}\hbar\partial/\partial t$ and $\boldsymbol{p}$ with $-\mathrm{i}\hbar\nabla$.

[137] Minimal coupling refers to a coupling between fields which involves only the charge distribution and not higher multipole moments of the charge distribution ([W11]).

[138] David Joseph Bohm (1917 – 1992) was an American physicist.

[139] Yakir Aharonov (1932 –) is an Israeli physicist.

[140] The Schrödinger equation is of the first order in time but of the second order with respect to $\boldsymbol{x}$. So it cannot be Lorentz covariant, because this transformation replaces t and $\boldsymbol{x}$ with linear combinations of t and $\boldsymbol{x}$.

[141] See Chapter 2.

sections 5.6 and 5.8 we shall therefore discuss two generalizations – from non-relativistic to relativistic for spin-0 particles, and from spin-0 to spin-½ ([A1]). (142)

5.3. The Schrödinger probability density current

In the previous section, we have seen that the (normalised) wave function ψ defines the probability density $\rho_S(t, \boldsymbol{x}) := |\psi(t, \boldsymbol{x})|^2 = \psi(t, \boldsymbol{x})\psi^*(t, \boldsymbol{x})$, where ψ^* denotes the complex conjugate of ψ.

In quantum mechanics probability is conserved, so if you observe that the probability of finding a particle at some point changes in time, then you deduce, that there must be a probability current flowing in or out. For example, in the case of a free particle solution, the probability density is uniform over all space, but there is a net flow along the direction of the momentum. However, the conservation law is local, so just like for the charge conservation (see Section 4.2). We may derive the conservation of the probability in Schrödinger's equation from the local continuity equation. Taking a time derivative of $\rho_S(t, \boldsymbol{x})$ and using the Schrodinger equation (5.4) for ψ and ψ^* we get

$$\partial\rho_S(t, \boldsymbol{x})/\partial t = \nabla \frac{1}{2mi}[\psi(t, \boldsymbol{x})\nabla\psi^*(t, \boldsymbol{x}) - \psi^*(t, \boldsymbol{x})\nabla\psi(t, \boldsymbol{x})].$$

The 3-vector

$$\boldsymbol{j}_S := \frac{1}{2mi}(\psi^*\nabla\psi - \psi\nabla\psi^*)$$

is called the *probability density current* of the wave function ψ. Rewriting the above equation we find the *continuity equation* for the probability density

$$\nabla\cdot\boldsymbol{j}_S = -\partial\rho_S/\partial t \tag{5.13}$$

which is analogous to the continuity equation for electric charge (4.2) and expresses mathematically the fact that probability density is locally conserved ([A1]). So the probability ρ_S in a tiny box decreases exactly by the amount that may be calculated as the flux of the probability current through the six faces of the little box (through its boundary) via Gauss' theorem. In particular, $\rho_S = |\psi|^2$ integrated over all space, is constant in time (143).

The quantity $(j_S)_\mu := (\rho_S, \boldsymbol{j}_S) = (|\psi|^2, \boldsymbol{j}_S)$ is a 4-vector being conserved 4-density current:

$$\nabla_\mu(j_S)^\mu = 0.$$

5.4. The gauge principle in electromagnetism

In the preceding section, we considered the Schrödinger equation (5.6) for a charged particle in an electromagnetic field. Then we showed its gauge covariance under the combined transformation (5.11').

---

142 It should be noted that to make quantum mechanics consistent with special relativity, the real problem is not to find a relativistic generalization of the Schrödinger equation. Quantum mechanics, as formulated by Bohr, Heisenberg, Schrödinger, Pauli, Dirac, and many others, is an intrinsically non-relativistic theory. Wave equations, relativistic or not, cannot account for processes in which the number and the type of particles changes, as in almost all reactions of nuclear and particle physics. This issue is solved with methods of quantum field theory (QFT) ([M1]).

143 Notice that this implies that the total probability to find a particle at any position is conserved. A drawback here is that (ordinary) QM cannot hope to describe a theory in which the number of particles changes with time. This is easy to see: if a particle disappears, then the total probability to find it beforehand should be unity and the total probability to find it afterwards should be zero. This issue is solved in quantum field theory. The QFT formalism allows considering a changing number of particles within the same framework. This requires the use of the concept of Fock space, and the use of creation and annihilation operators. The entire process of setting this up is called *second quantisation* ([G7]). A Fock space is a special construction of a Hilbert space (cf. Section 5.9). The basic idea is that the Fock space allows you to superpose tensor products of distinct degree. Thus one can describe states on which the very number of particles is uncertain and becomes an observable with probabilities and mean values as any other observable (see Section 5.9 for the meaning of observables and states in quantum physics).

Now let us start with the free-particle Schrödinger equation (5.4) and discuss consequences of demanding covariance under the spacetime-dependent (i.e. local) transformation of the wave function ([A1])

$$\psi(t, \boldsymbol{x}) \rightarrow \psi'(t, \boldsymbol{x}) = e^{iq\Lambda(t,\boldsymbol{x})}\psi(t, \boldsymbol{x}). \tag{5.14}$$

The situation in which the wave function can be changed in a certain way without leading to any observable effects is precisely what is entailed by a symmetry or invariance principle in quantum mechanics. In the case of a constant overall change in phase $\psi \rightarrow \psi' = e^{i\alpha}\psi$, $\alpha =$ constant, the invariance principle guarantees that any choice of α is equivalent to any other. This symmetry is an internal symmetry, because it is clearly no spacetime transformation and therefore transforms the field ψ internally. This internal symmetry does not look at a first glance like a big thing. However, quite surprisingly we shall see in a moment that this symmetry is incredibly important when it is made local. ([A1])

Invariance under a constant change in phase is an example of a global invariance, according to the terminology introduced in Section 3.5. However, as we have seen in Section 4.4, the interconnection of electric and magnetic effects can be related to a *local* invariance requirement − in this case, electromagnetic gauge invariance. Such a move from a global to a local invariance is of crucial significance in classical electromagnetism and provides also the key to an understanding of the other interactions in the Standard Model. ([A1])

Let us see, then, where the demand of *local phase invariance* (5.14) of the field ψ leads us (or rather, more accurately, a corresponding covariance). The immediate problem is that this is obviously not a covariance of the free-particle Schrödinger equation (5.4). The equation (5.4) does not have the same form when we replace in it $\psi(t, \boldsymbol{x})$ by $e^{iq\Lambda(t,\boldsymbol{x})}\psi(t, \boldsymbol{x})$. The reason is that both $\boldsymbol{\nabla}$ and $\partial/\partial t$ now act on $\Lambda(t, \boldsymbol{x})$ in the phase factor. Thus local phase covariance is not a covariance of the free-particle wave equation (5.4). ([A1])

Consequently, if we demand this covariance, we have to modify the equation (5.4) into something for which there is a local phase covariance. But this modified equation will no longer describe a free particle. Thus the freedom to alter the phase of a charged particle's wave function locally is only possible if some kind of force field is introduced in which the particle moves. In more physical terms, the covariance will now be manifested in the inability to distinguish observationally between the effect of making a local change in phase and the effect of some new field in which the particle moves. ([A1])

What kind of field will this be? Since the local phase transformation (5.14) is just the phase transformation associated with electromagnetic gauge invariance (5.11), we must modify the free-particle equation $i(\partial/\partial t)\psi = -\frac{1}{2m}\nabla^2\psi$ to

$$i(\partial/\partial t + iqV)\psi = -\frac{1}{2m}(\boldsymbol{\nabla} - iq\boldsymbol{A})^2\psi,$$

which is precisely the Schrödinger equation (5.6) describing the interaction of the charged particle with the electromagnetic field $A_\mu = (V, \boldsymbol{A})$. Thus the presence of the vector field A_μ, interacting with any particle of charge q, is dictated by local phase invariance. ([A1])

Such a vector field, introduced to guarantee local phase invariance, is called a *gauge field*. The principle that the interaction should be so dictated by the local phase (or gauge) invariance is called the *gauge principle*. This principle allows one to write down the wave equation for the interaction directly from the free particle equation via replacement ([144]) ([A1])

$$\nabla_\mu \rightarrow D_\mu = \nabla_\mu + iqA_\mu.$$

[144] It is important to note that the form of the covariant derivative depends on the transformation properties of the field on which it acts. For instance, for fields transforming as in Eq. (5.14), D_μ depends on the parameter q ([M1]).

This procedure is expressed by saying that we have *gauged* the global U(1) symmetry, promoting it to a local symmetry. The resulting theory is called a *gauge theory*. More precisely, it is a U(1), or Abelian gauge theory, since we have gauged a U(1) symmetry ([M1]) – cf. Section 5.5.

To summarise:

- Consider a matter system with global gauge invariance, which guarantees the existence of a conserved charge.
- The global gauge invariance is then to be extended to local gauge invariance through the replacement $\nabla_\mu \to D_\mu$, thereby introducing a coupling to a gauge field.

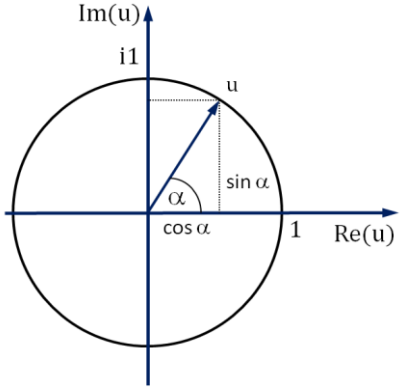
As we shall see in Chapter 6, the gauge principle led to the discovery of Yang-Mills theory, which is a non-Abelian gauge theory ([145]).

Finally, let us observe that the gauge invariance requires that photon as the boson of the gauge field A_μ must be massless. Otherwise, we would have to add to the Lagrangian density a non-invariant term $\mathcal{L}_{mass} = \frac{1}{2} m^2 A^\mu A_\mu$ (= the rest-mass energy of photon) ([146]). And since this is true for any gauge field, it has important implications for every Yang-Mills theory.

5.5. The gauge group U(1) of electromagnetism

Let us look more closely at the nature of the symmetry related to invariance under (5.14). It is not a symmetry which – as in the case of Lorentz invariances for instance – involves changes in the spacetime coordinates. Instead, it operates on the real and imaginary parts of the field ψ.

We shall see that the transformation (5.14) is a kind of rotation in two dimensions. There is a way to describe rotations in two dimensions that makes use of complex numbers: rotations about the origin by angle α can be described by multiplication with a unit complex number $u = v + iw$ which fulfils the condition $|u|^2 = u^*u = 1$, where $u^* = v - iw$ denotes the complex conjugate of u. For a complex number $z = x + iy$, x is called the real part of z: $\mathrm{Re}(z) = x$ and y the imaginary part: $\mathrm{Im}(z) = y$. The unit complex numbers lie on the unit circle in the complex plane:



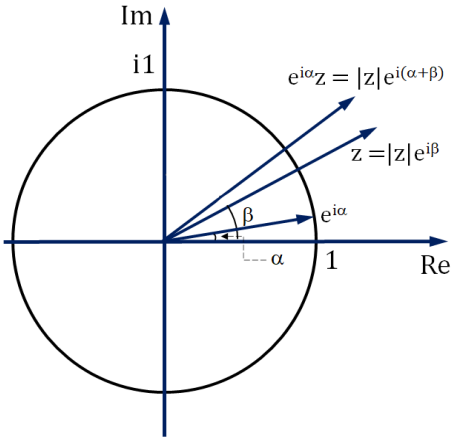
Another way to write a complex number $z = x + iy$ is $z = |z|e^{i\beta} = |z|(\cos\beta + i\sin\beta)$ ([147]). Thus for a unit complex number $u = e^{i\alpha}$ the transformation $z \to uz = e^{i\alpha}z = e^{i\alpha}|z|e^{i\beta} = |z|e^{i(\alpha+\beta)} =$ means that z is rotated by the angle α in the Re-Im plane

---

[145] Yang and Mills replaced the gauge group U(1) of electromagnetism by the non-Abelian group SU(2) – see Chapter 6 for details.

[146] The term $A^\mu A_\mu$ is not invariant under the gauge transformation $A_\mu \to A_\mu - \nabla_\mu\Lambda$ because $(A^\mu - \nabla^\mu\Lambda)(A_\mu - \nabla_\mu\Lambda) \neq A^\mu A_\mu$. Reminder: $A^\mu := \eta^{\nu\mu}A_\mu = (V, -\boldsymbol{A})$, where $\eta^{\nu\mu}$ is the Minkowski metric.

[147] This is known as Euler's formula.



We call this plane an *internal space* and the associated symmetry an *internal symmetry*. Thus the phase invariance (5.14) can be looked upon as a kind of internal space rotational invariance.

The set of all phase transformations $z \to u_\alpha(z) := e^{i\alpha}z$, with real α, forms a group called U(1) ([148]). It is an Abelian group since the transformations of the U(1) have the simple property that it does not matter in what order they are performed: $z \to (u_\alpha u_\beta)(z) = e^{i(\alpha+\beta)} = e^{i(\beta+\alpha)} = (u_\beta u_\alpha)(z)$.

Thus we say that the electromagnetic *gauge group* is U(1). We must remember, however, that it is a local U(1), signifying that the phase parameters α, β, . . . depend on the spacetime point $(t, \boldsymbol{x})$. In other words, this means that an independent U(1) gauge group is to be associated with each spacetime point. To achieve this, it is necessary to introduce a vector gauge field, to which the matter field current becomes coupled. The coupling constant is the electric charge, the generator of U(1) (see Example 7.1). It is important to note that the original global symmetry can be gauged only if it is an exact symmetry.

We shall see later (in Chapter 6) that the 'internal' symmetry space relevant to the Yang-Mills invariance is not so simple. The 'rotations' in this case are more like full three-dimensional rotations of real space, rather than the above two-dimensional rotations in the Re-Im plane. However in general, such 3D real-space rotations do not commute (see Example 3.3), and the same will be true of the Yang-Mills 'rotations' which build a non-Abelian group. The generalized gauge principle, as initially proposed by Yang and Mills, applies to multicomponent matter fields. Thus instead of U(1), the gauge group must be a larger group of transformations that mix the different components of the matter field ([H14]).

5.6. The Klein-Gordon equation

Physical equations should be Lorentz covariant, meaning that they must be covariant under Lorentz transformations – that is, they must have the same form in the two different reference frames. The simplest relativistic wave equation is the Klein-Gordon ([149]) equation which describes a scalar particle (i.e. spin-0 particle) of mass m ([150]).

For a relativistic wave equation we must start with the correct relativistic energy–momentum relation. Energy and momentum appear as the 'time' and 'space' components of the momentum 4-vector $p^\mu = (E/c, \boldsymbol{p}) = (E, \boldsymbol{p})$ which satisfy the condition $p^2 = p_\mu p^\mu = E^2 - \boldsymbol{p}^2 = m^2$ − see Section 3.3 ([151]). Schrödinger, before settling for the less ambitious non-relativistic

---

[148] The group operation $(u_\alpha, u_\beta) \to u_\alpha u_\beta$ on U(1) is the usual function composition $z \to (u_\alpha u_\beta)(z) = u_\alpha(u_\beta(z))$.

[149] Oskar Benjamin Klein (1894 –1977) was a Swedish theoretical physicist; Walter Gordon (1893 – 1939) was a German theoretical physicist.

[150] Actually, Schrödinger first derived a relativistic equation, that today we call the Klein-Gordon equation. He then discarded it because it gave the wrong fine structure for the hydrogen atom, and he retained only the non-relativistic limit. His relativistic wave equation was independently rediscovered by O. Klein and W. Gordon ([M1], [W1]).

[151] Reminder: $c = \hbar = 1$.

Schrödinger equation, and later Klein and Gordon, attempted to build relativistic quantum mechanics (RQM) from the squared relation $E^2 = \boldsymbol{p}^2c^2 + m^2c^4$ instead of from the non-relativistic energy–momentum relation $E = \frac{\boldsymbol{p}^2}{2m}$. Using the operator replacements $E \to \mathrm{i}\hbar\partial/\partial t$ and $\boldsymbol{p} \to -\mathrm{i}\hbar\nabla$ we obtain ([A1])

$$\hbar^2\partial^2/\partial t^2\psi = (\hbar^2\mathrm{c}^2\nabla^2 - m^2c^4)\psi \tag{5.15}$$

which is the *Klein–Gordon equation* (KG equation). We consider the case of a one-component scalar (complex) wave function $(\mathrm{t}, \boldsymbol{x}) \to \psi(\mathrm{t}, \boldsymbol{x}) \in \mathbb{C}$: one expects this to be appropriate for the description of spin-0 bosons. Applying the d'Alembertian $\Box^2 = \nabla_\mu\nabla^\mu = \partial^2/\partial t^2 - \nabla^2$ we can, using natural units, write the Klein-Gordon equation in the form

$$(\Box^2 + m^2)\psi = 0. \tag{5.16}$$

Now let us ask if there is a Lagrangian density for the (free) field ψ from which we can derive the Klein-Gordon equation (KG) by the principle of least action. Let us assume that someone hands us the Lagrangian density ([152])

$$\mathcal{L}_{KG}^{free} = \frac{1}{2}(\nabla_\mu\psi^*\nabla^\mu\psi - m^2\psi^2),$$

where ψ^* denotes the complex conjugate of ψ. It is conventional to call the first term here the *kinetic term* and the second one the *mass term*. Applying the Euler-Lagrange equation to this Lagrangian one can easily derive the corresponding equation of motion which is exactly (5.15).

If a (spin-0) particle has electric charge q (like meson π^+ or π^-), one needs to suitably insert the electromagnetic potential A_μ into the Klein-Gordon equation. More precisely, the equation

$$(\Box^2 + m^2)\psi = -\mathrm{i}q[\nabla_\mu A^\mu + A_\mu\nabla^\mu]\psi + q^2A^\mu A_\mu\psi \tag{5.17}$$

describes a scalar particle of mass m and charge q in the presence of an electromagnetic field.

Consider a Lorentz transformation such that $\mathrm{x} \to \mathrm{x}'$, i.e. $(\mathrm{t}, \boldsymbol{x}) \to (\mathrm{t}', \boldsymbol{x}')$ and write the transform of ψ as $\psi(\mathrm{t}, \boldsymbol{x}) \to \psi'(\mathrm{t}', \boldsymbol{x}')$. Since x' is a known function of x, given by the angles and velocities parameterising the Lorentz transformation (see Section 3.2), one can construct the correct function ψ' which the primed observers must use, in order to be consistent with the unprimed observers. Consequently, the wave function in the primed frame may be identified (up to a phase) with that in the unprimed frame: $\psi'(\mathrm{t}', \boldsymbol{x}') = \psi(\mathrm{t}, \boldsymbol{x})$. Now the 4-dimensional dot products appearing in (5.17) are all invariant under the Lorentz transformation so that equation (5.17) is covariant under Lorentz transformations. ([A1])

Notice that we can derive the KG equation (5.17) from the free equation (5.16) by applying the gauge principle. All we have to do is to replace ∇_μ by $D_\mu = \nabla_\mu + \mathrm{i}qA_\mu$ (see Section 5.4).

Indeed, let us replace ∇_μ by D_μ in $(\Box^2 + m^2)\psi = 0$:

$$\begin{aligned}
0 &= (\Box^2 + m^2)\psi \\
&= (\nabla_\mu\nabla^\mu + m^2)\psi \\
&= (D_\mu D^\mu + m^2)\psi \\
&= (\nabla_\mu + \mathrm{i}qA_\mu)(\nabla^\mu + \mathrm{i}qA^\mu)\psi + m^2\psi \\
&= (\nabla_\mu\nabla^\mu + \mathrm{i}q\nabla_\mu A^\mu + \mathrm{i}qA_\mu\nabla^\mu - q^2A_\mu A^\mu)\psi + m^2\psi \\
&= (\nabla_\mu\nabla^\mu + m^2)\psi + \mathrm{i}q[\nabla_\mu A^\mu + A_\mu\nabla^\mu]\psi - q^2A_\mu A^\mu\psi \\
&= (\Box^2 + m^2)\psi + \mathrm{i}q[\nabla_\mu A^\mu + A_\mu\nabla^\mu]\psi - q^2A_\mu A^\mu\psi.
\end{aligned}$$

Consequently, we obtain $(\Box^2 + m^2)\psi = -\mathrm{i}q[\nabla_\mu A^\mu + A_\mu\nabla^\mu]\psi + q^2A_\mu A^\mu\psi$ as needed.

[152] This choice can be motivated by looking at the Lagrangian for a classical harmonic oscillator ([A3]).

Since the KG equation is derived from the squared relation $E^2 = \boldsymbol{p}^2 + m^2$, it implies that for a given 3-momentum $\boldsymbol{p}$ there are in fact two possible solutions for the energy, namely

$$E = \pm(\boldsymbol{p}^2 + m^2)^{1/2}.$$

As Schrödinger and others quickly found, it is not possible to ignore the negative solutions without obtaining inconsistencies. Negative energy states are problematic because there is nothing to stop the vacuum decaying into these states ([153]). In classical relativistic mechanics, the problem of these negative energy solutions never appears, because we could simply throw them away, declaring that all particles (or rockets or whatever) have positive energy. But when we solve a wave equation (as we do in QM), completeness requires us to include both positive and negative energy solutions in order to be able to find a general solution. ([A1])

But there is one more problem with the KG equation. In exactly the same way as for the non-relativistic Schrödinger equation (cf. (5.13)), it is possible to derive a conservation law for a 'probability current' of the Klein-Gordon equation

$$\nabla \cdot \boldsymbol{j}_{KG} = -\partial \rho_{KG}/\partial t,$$

where

$$\rho_{KG} := \mathrm{i}(\psi^* \partial\psi/\partial t - \psi \partial\psi^*/\partial t)$$

and

$$\boldsymbol{j}_{KG} := \frac{1}{\mathrm{i}}(\psi^* \nabla\psi - \psi \nabla\psi^*).$$

So far so good, but note that the 'probability density' ρ_{KG} now contains time derivatives. This means that ρ_{KG} is not constrained to be positive definite ([154]) – so how can ρ_{KG} represent a probability density? ([A1]) ([155])

Historically, this problem of negative probabilities coupled with that of negative energies led to the abandonment of the Klein-Gordon equation. However, these issues were later on solved within the formalism of quantum field theory. ([A1])

5.7. An interlude on spinors, helicity and chirality ([156])

Spinors began to find a more extensive role in physics when it was discovered that fermions have a half-integral spin which is correctly captured by the mathematics of spinors. Pauli ([157])

---

[153] Energy spectrum is in this case not bounded from below, so a particle can emit an infinite amount of energy (no ground state). QFT solves this problem by interpreting negative solutions as antiparticles.

[154] The KG equation has a $\partial^2/\partial t^2$ term: this leads to a 'probability density' containing $\partial/\partial t$, and hence to negative probabilities.

[155] In general, a physical theory for calculating probabilities should not be dismissed as erroneous if it yields a negative probability for a given situation under specific assumed conditions. There may be alternative explanations that could be postulated ([F12]).

See also [K0] for p-adic probability theory. In the field of p-adic probability theory, probabilities are expressed through the utilisation of p-adic numbers as opposed to real numbers. This approach facilitates the exploration of novel mathematical structures and models, particularly within the domain of quantum mechanics. Notably, within the framework of p-adic probability theory, probabilities can assume negative values. The theory utilises an alternative metric, termed the p-adic metric, to define probabilities. This enables the consistent and mathematically rigorous incorporation of negative probabilities.

[156] For further information, please refer to the comprehensive introduction to spinors [S12] and the detailed analysis of helicity and chirality in [S0], respectively.

[157] Wolfgang Pauli (1900 – 1958) was an Austrian (and later American / Swiss) theoretical physicist and one of the pioneers of quantum physics – 1945 Nobel Prize in Physics.

modelled in 1927 the electron spin in a non-relativistic context using a two-component complex vector and introducing the Pauli spin matrices (5.23). ([158])

We shall see in the next section that spinors are important for the Dirac equation which describes all of the known fundamental particle fermions in nature. All known fermions, the particles that constitute ordinary matter, have a spin of ½ (see Chapter 2). The spin number describes how many symmetrical facets a particle has in one full rotation; a spin of ½ means that the particle must be fully rotated twice (through 720°) before it has the same configuration as when it started. Ordinarily, when one rotates an object 360°, it goes back to the same thing it started out as. This is common sense, but as we learned in quantum mechanics, common sense can be misleading. The necessity of introducing half-integer spin goes back experimentally to the results of the Stern–Gerlach experiment in 1922 ([159]).

A spinor is a mathematical object that models this strange behaviour of spin-½ particles. It appears that Felix Klein ([160]) originally designed the spinor to simplify the treatment of the classical spinning top in 1897. As mathematical objects, spinors were introduced in geometry by Élie Cartan ([161]) in 1913. They are closely related to Hamilton's quaternions (about 1845).

Cartan defined spinors in terms of isotropic 3-vectors with complex components: $\boldsymbol{z} = (z_1, z_2, z_3) \in \mathbb{C}^3$. A vector $\boldsymbol{z}$ is said to be *isotropic* if its dot (scalar) product with itself is zero:

$$\boldsymbol{zz} := z_1^2 + z_2^2 + z_3^2 = 0.$$

---

[158] In 1921, Arthur Compton (an American physicist, Nobel Prize in Physics in 1927) suggested that electron spin would be an essential ingredient in any reasonable explanation of bulk paramagnetism and ferromagnetism. Unfortunately, Compton's proposal had almost no impact on his contemporaries. At that time, the anomalous Zeeman effect was a persistent puzzle that absorbed the attention of several physicists. Anomalous Zeeman effect is the splitting of spectral lines of an atomic spectrum caused by the interaction between magnetic field, the combined orbital and intrinsic magnetic moment. This effect can be observed as a complex splitting of spectral lines.

On January 7, 1925, the 20-year-old German Ralph Kronig (a Columbia University PhD student), when visiting Tübingen, was shown a letter from Pauli (who was 25 years old) in which Pauli emphasized that to understand the anomalous Zeeman effect, it would be necessary to endow the electron with a fourth quantum number with only two discrete values. On reading this letter, Kronig was struck with inspiration, and that very afternoon he invented a concept of electron spin. His mental picture of the electron was of a tiny spinning classical sphere, and his interpretation of Pauli's fourth quantum number was that the spin axis could point in only two (opposing) directions. The next day Kronig explained his idea to Pauli, who said: "*it is indeed very clever but of course has nothing to do with reality*". Pauli realized that the fourth quantum number corresponded to a classically non-describable degree of freedom, so he must have viewed Kronig's classical picture of the electron as unacceptably naïve. Several weeks later Kronig presented his idea again at Niels Bohr's institute in Copenhagen. Unfortunately, Bohr and others also gave it a cold reception, objecting on the same grounds as Pauli had. Faced with such criticism, Kronig decided not to publish his theory and the idea of electron spin had to wait for others to take the credit.

The next development occurred in autumn 1925 in Leiden, Holland, where Uhlenbeck, 24, and Goudsmit, 23, were students of the professor of theoretical physics Paul Ehrenfest. Uhlenbeck and Goudsmit, unaware of Kronig's efforts, essentially reinvented the latter's idea in one afternoon. They had a much more kind review on their work from Ehrenfest, who said that it was either nonsense or something very important, that they should write up a short paper, and that all three would then consult Professor Hendrik Lorentz. Lorentz listened courteously to them and promised to give them his reply. In several days, Lorentz did reply at length in a handwritten manuscript in which he gave a number of serious objections to Uhlenbeck and Goudsmit's proposal. They told Ehrenfest of this and wished to withdraw their paper. However, Ehrenfest had already sent it to the publisher. He told them not to worry: they were young enough to be forgiven for their stupidity.

In February 1926, Llewellyn H. Thomas, a 23-year-old British physicist, realized that the spin-orbit effect described by Uhlenbeck and Goudsmit required a relativistic correction. He calculated relativistic effects on the spin–orbit interaction in a hydrogen atom (Thomas precession) and could understand the anomalous Zeeman effect. In March 1926, when Pauli learned of Thomas's result, he was converted to the idea of electron spin, and within about a year he developed a formalism for describing the spinning electron in non-relativistic quantum mechanics – see [C5] for a comprehensive history of spin discovery which is well worth reading .

[159] Otto Stern (1888 – 1969) was a German-American physicist and Nobel laureate in Physics 1944. Walther Gerlach (1889 – 1979) was a German physicist. The experiment was conceived by Stern in 1921 and first successfully conducted by Gerlach in early 1922.

[160] Felix Klein (1849 – 1925) was a German mathematician (known for the Erlangen program and the Klein bottle).

[161] Élie Cartan (1869 – 1951) was an influential French mathematician.

For example

$$\boldsymbol{z} = (1, \mathrm{i}, 0) = (1+\mathrm{i}0, 0+\mathrm{i}, 0+\mathrm{i}0)$$

is isotropic because $\boldsymbol{zz} = 1 + (-1) + 0 = 0$.

To every isotropic 3-vector $\boldsymbol{z} = (z_1, z_2, z_3)$ one can assign a complex 2-vector $\phi = (\phi_1, \phi_2) \in \mathbb{C}^2$ whose components satisfy the equations:

$$\phi_1^2 - \phi_2^2 = \mathrm{z}_1, \quad \mathrm{i}(\phi_1^2 + \phi_2^2) = \mathrm{z}_2, \quad -2\phi_1\phi_2 = \mathrm{z}_3. \tag{5.18}$$

Notice that if ϕ solves the system (5.18) then $-\phi$ is also its solution. Thus there are two 2-vectors corresponding to an isotropic 3-vector. These 2-vectors are called *spinors*. One can show that

$$\phi_1 = \pm\sqrt{\frac{z_1 - iz_2}{2}}, \quad \phi_2 = \pm\sqrt{\frac{-z_1 - iz_2}{2}}.$$

However, spinors and vectors are completely different kinds of objects. To see it, let us consider a rotation by 360° around an arbitrary axis. A vector remains completely unchanged by such a full rotation. A spinor on the other hand is not unchanged by a full rotation – it changes its sign. We need to rotate a spinor twice by 360° (i.e. by 720°) to get it back to its initial configuration.

Example 5.1. Let us consider the isotropic 3-vector $\boldsymbol{z} = (1, \mathrm{i}, 0)$. One corresponding spinor is then $\phi = (1, 0)$. If $\boldsymbol{z}$ is rotated in $\mathbb{C}^3$ around the z-axis by the angle θ we get the 3-vector

$$\begin{aligned}\boldsymbol{z}' &= (\cos(\theta) - \mathrm{i}\sin(\theta), \quad \sin(\theta) + \mathrm{i}\cos(\theta), \quad 0)\\ &= (\cos(-\theta) + \mathrm{i}\sin(-\theta), \mathrm{i}(\cos(-\theta) + \mathrm{i}\sin(-\theta)), 0)\\ &= (e^{-\mathrm{i}\theta}, \mathrm{i}e^{-\mathrm{i}\theta}, 0) = e^{-\mathrm{i}\theta}(1, \mathrm{i}, 0) = e^{-\mathrm{i}\theta}\boldsymbol{z}.\end{aligned}$$

Now let us look at the spinor ϕ' corresponding to $\boldsymbol{z}'$

$$\phi' = \left(\sqrt{\frac{e^{-\mathrm{i}\theta} - \mathrm{i}(\mathrm{i}e^{-\mathrm{i}\theta})}{2}}, \sqrt{\frac{-e^{-\mathrm{i}\theta} - \mathrm{i}(\mathrm{i}e^{-\mathrm{i}\theta})}{2}}\right) = \left(\sqrt{\frac{2e^{-\mathrm{i}\theta}}{2}}, 0\right) = (e^{-\mathrm{i}\theta/2}, 0) = e^{-\mathrm{i}\theta/2}\phi.$$

But it means a rotation of ϕ around the first axis by the angle $\theta/2$.

Generally, the 2×2 matrix that describes how a spinor transforms under rotation around the first axis is

$$R(\theta) = \begin{bmatrix} \cos\frac{\theta}{2} & i\sin\frac{\theta}{2} \\ i\sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{bmatrix}.$$

For $\theta = 2\pi$ we get

$$R(2\pi) = \begin{bmatrix} \cos\frac{2\pi}{2} & i\sin\frac{2\pi}{2} \\ i\sin\frac{2\pi}{2} & \cos\frac{2\pi}{2} \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} = -\mathrm{I}_2.$$

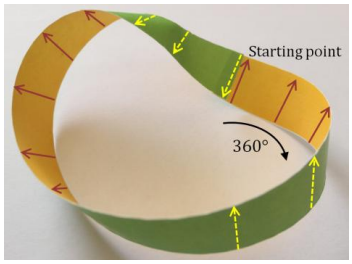
This means that a spinor changes the sign after a full rotation. Notice that $R(4\pi) = \mathrm{I}_2$, where I_2 denotes the 2×2 identity matrix. So we need to rotate a spinor by 4π (i.e. by 720°) to get it back to its initial configuration ([S2]).

There are several ways of illustrating spinors using everyday analogies in terms of the Dirac's belt, tangloids and other examples of orientation entanglement. Nonetheless, the concept is generally considered notoriously difficult to understand ([162]). The Dirac's belt is related to a

---

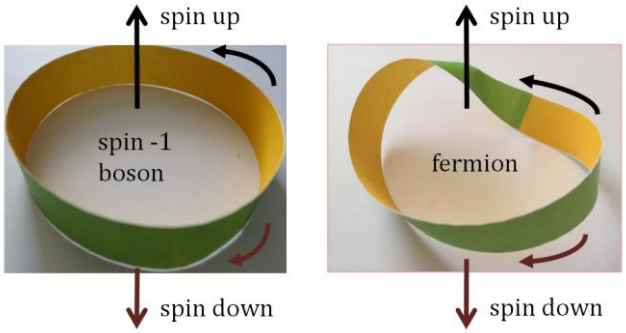
[162] Even M. Atiyach (winner of the Fields Medal) declared that "*no one fully understands spinors. Their algebra is formally understood, but their geometrical significance is mysterious. In some sense they describe the 'square root' of geometry and, just as understanding the concept of* $\sqrt{-1}$ *took centuries, the same might be true of spinors.*" ([S2]). R. Feynman was once asked by a Caltech faculty member (David L. Goldstein) to explain why spin one-half particles obey Fermi-Dirac statistics. Feynman said, "*I'll prepare a freshman lecture on it.*" But a few days later he told the faculty member: "*You know, I couldn't do it. I couldn't reduce it to the freshman level. That means we really don't understand it.*" ([G4] I thank Prof. M. Bulenda for pointing to this reference.) Feynman meant here that understanding something is not just about working through advanced mathematics. One must also have a notion that is intuitive enough to explain to an audience that cannot follow the detailed derivation.

Möbius strip. A spinor can be visualized as a vector pointing along the Möbius strip, exhibiting a sign inversion when the circle (the 'physical system') is continuously rotated through a full turn of 360° ([W11]):



One has to go around one more time to come back to your original orientation. Spinors have the same property. That is why a spinor can be illustrated by a Möbius strip. One gets this strip by cutting a regular strip in one spot, turn one end by 180° and connect it back to the other end.

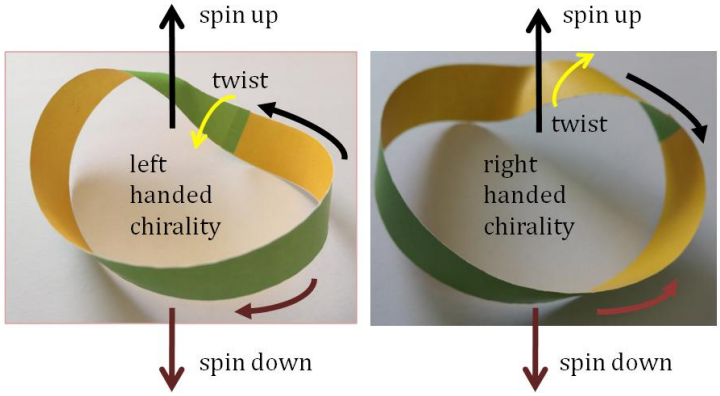
Following [S0], we can view spin as a wave going around on a circle. The spin wave of a (vector) boson goes around on a regular circle and is quantized into integer units. The spin wave of a fermion (spin-½ particle) goes around a Möbius strip and takes two rounds to get back to its original position. Thus we can connect spin of bosons and fermions to the two types of rotation depicted below



Notice that there are two ways of making a Möbius strip. You can turn one end of a regular strip by 180° in two directions before you tape it back to the other end. When you turn the one end by 180° left around you get one chirality of the Möbius strip. If you turn it 180° right around you get its mirror image: the other chirality. The mirror image of a Möbius strip is again a Möbius strip, but one that cannot be turned in three dimensions to be equal to its original.

Chirality is the property of an object that it has a mirror image asymmetric to itself. For instance a left hand. In the mirror a left hand looks like a right hand. And you cannot rotate your right hand in any way so it becomes your left hand. Both 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. This can be done in each chirality in an equal manner.

The combinations chirality/spin correspond exactly with the four ways a fermion can rotate, as depicted below



A spin wave in one chirality goes through a different path with different directions of the spinor field compared to the other chirality. The consequence is that we have to distinguish the two chiralities as two different particle types: *left-chiral* and *right-chiral* ([163]). It makes the behaviour of fermions more complex as we shall see in Section 7.4. One interesting problem is, for example, that a neutrino exists only as a left-chiral particle. It is unknown why there is no right-chiral neutrino. This too suggests considering the left and right-chiral particles as a different breed ([S0]).

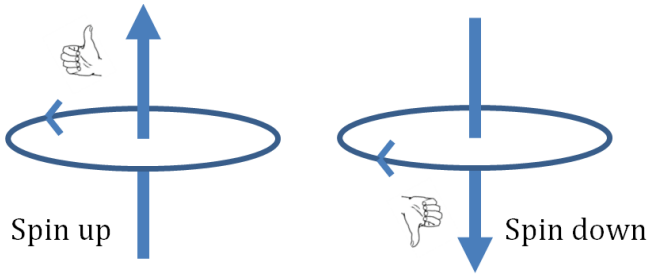
In conclusion, the following four possible fermion states can be identified:

- left chirality and spin-(+½)
- left chirality and spin-(−½)
- right chirality and spin-(+½)
- right chirality and spin-(−½).

It is important to note, however, that the above visualisation utilising the Möbius strip is merely a picture that assists in comprehending the underlying principles, yet does not necessarily reflect the absolute truth of what spin is. What spin really is, is not known. We can only measure the characteristics, and we cannot see directly what spin is ([S0]).

Both spin and chirality are related to *helicity*, which is also a property of particles. However, it is noteworthy that the terms *spin*, *helicity*, *chirality*, and *handedness* are not used consistently in the literature. The issue is that the term 'handedness' is a highly intuitive and convenient way to describe both chirality and helicity. Consequently, courses and textbooks may employ the term 'handedness' to describe the term that is used more frequently. There is no established standard. In order to facilitate clarity, a brief clarification of their precise relationship is provided.

The direction of spin is described by an arrow positioned perpendicular to the surface of the spin circle. The direction of the spin arrow is determined by the right-hand rule. It is evident that there are merely two potential orientations for rotation: left around, formally designated as *spin up*, and right around, formally designated as *spin down*.



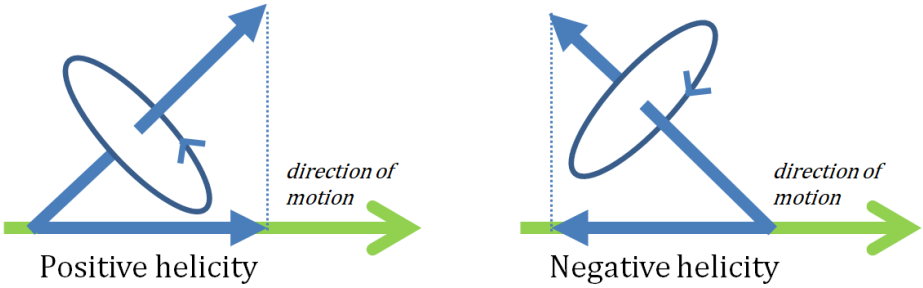
The entire concept is based on convention. There is no underlying significance to the right hand rule.

The term *helicity* is used to describe the component of spin that is aligned with the direction of the momentum vector. In other words, helicity is the projection of the spin vector upon the momentum. Given the existence of two distinct spinning directions, helicity is also observed to assume two distinct values. These are as follows:

1. *positive helicity*, represented by +1, whereby the spin is oriented in the direction of propagation; and

2. *negative helicity*, represented by −1, whereby the spin is oriented in the opposite direction.

[163] Also called left-handed chirality and right-handed chirality, respectively.



It hence follows that a particle can possess a helicity of either $+1$ or -1, provided it is not a scalar ([164]) ([S0]). It is also possible for massive non-scalar particles to possess neutral helicity, represented by the value of 0, when the spin is perpendicular to the direction of motion.

The helicity of a massive particle with spin is not Lorentz invariant. Indeed, in this case it is possible for an observer to change to a reference frame moving faster than the spinning particle. In such a frame, the particle will appear to move backwards, and its helicity will be reversed. In other words, the helicity may change sign under the action of a Lorentz boost.

In contrast, a massless particle moves at the speed of light, so a real observer (who must always travel at less than the speed of light) cannot be in any reference frame where the particle appears to reverse its relative direction. Consequently, all observers see the same helicity. As result, the direction of spin of massless particles is not affected by a Lorentz boost in the direction of their motion, and the sign of the projection (helicity) is constant across all reference frames. This demonstrates that the helicity of a massless particle is a relativistic invariant.

We will now proceed to discuss the relationship between helicity and chirality.

The term *chirality* is used to describe the property of an object that possesses a mirror image which is asymmetric to itself. In the preceding discussion, the term chirality was defined as the twist direction of the Möbius strip ([165]). They are each other's mirror images. The fundamental distinction between bosons and fermions is their spin. As previously indicated, fermions manifest a spin wave around a Möbius strip, whereas bosons display a spin wave around a regular circle. Consequently, fermions are classified into two chiralities, whereas this property is not exhibited by bosons. Both chiralities have identical effects; a circuit around one leads to the other side of the Möbius strip. It can be concluded that the chirality of the Möbius strip has no impact on the spin waves. In both chiralities, the spin can be either up or down. ([S0])

In the case of massive fermions it is necessary to distinguish between chirality and helicity. In this context, helicity is not a Lorentz-invariant attribute, while chirality possesses this property. However, chirality is not a constant of motion. A massive left-chiral fermion will evolve into a right-chiral fermion over time, and vice versa ([166]). Thus, its quantum mechanical state is a mixture of two opposing chiralities. ([W11])

The chirality remains unaltered when observed from different frames, as this would necessitate a parity operation that encompasses a mirror reflection. It is not possible to achieve

---

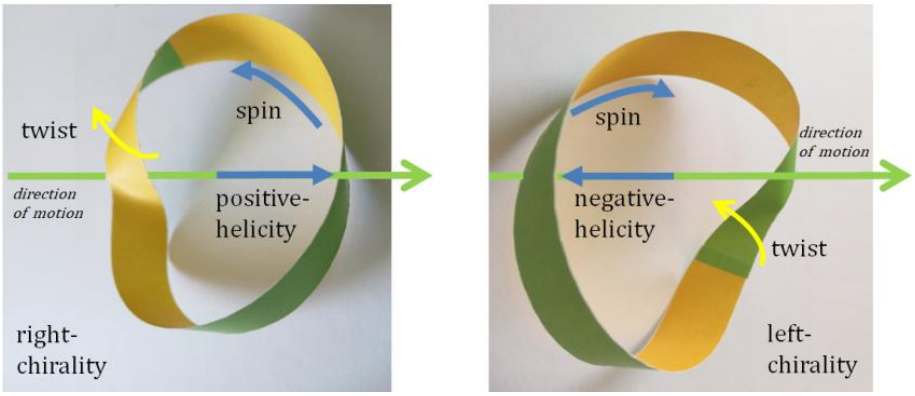
[164] The positive helicity of a particle is sometimes referred to as 'right-handed', while the negative helicity is designated as 'left-handed'. This may be confusing since the term 'handedness' is also used to describe chirality.

[165] Chirality is a concept that only exists for spinors. More precisely, it only exists for representations (s_1, s_2) of the Lorentz group with $s_1 \neq s_2$ (c.f. Example 7.5). In mathematical terms, chirality can be described as an eigenstate of the γ_5 matrix. Please refer to Section 5.8 for a more detailed explanation.

[166] Massive Dirac particles exhibit a coupling between left-chiral and right-chiral wave functions, which is determined by the mass associated with the particles in question. The implication is that such particles oscillate between left- and right-chiral states at a rate dependent on their mass. This oscillation is caused by interaction with the Higgs field. There is a correlation between the frequency of interaction with the Higgs field and the mass of the fermion; the higher the mass, the greater the frequency of Higgs field interaction, and the more frequent is the switching of chirality ([L1], [S0]). Please refer to section 5.8 for further details.

a mirror reflection by changing a frame of reference, which merely entails a different velocity and direction of motion. It is evident that helicity and chirality are two distinct concepts.

Nevertheless, at exceedingly high speeds (i.e. in the relativistic limit $E \gg m$), a right-chiral fermion is associated with a positive helicity (spin aligned with the direction of propagation), while a left-chiral fermion is associated with a negative helicity at the same velocity. Consequently, a massless fermion with spin up in the direction of motion has a fixed positive helicity and a fixed right-chirality. Similarly, a massless fermion with spin down in the direction of motion has a fixed negative helicity and a fixed left-chirality. This implies that there are only two permitted configurations of helicity and chirality ([S0]):



In other words, right-chirality is not possible in a negative helicity situation and left-chirality is not possible in a positive helicity situation.

Summarised: helicity and chirality are equivalent ([167]) for massless fermions (or in the relativistic limit $E \gg m$), but are, in general, not the same. ([168]).

5.8. The Dirac equation ([169])

In 1928, P.A.M. Dirac tried to solve the problem of negative-energy solutions by looking for a wave equation that was first order in time-derivatives, the hope being that one could then obtain a relation of the form $E = +(\boldsymbol{p}^2 + m^2)^{1/2}$ directly, without encountering negative energy states ([170]). Dirac realized that one could write an equation that was linear in both time and space derivatives of the form ([A1])

$$\begin{aligned} \mathrm{i}\partial\psi/\partial t &= [-\mathrm{i}(\alpha_1\partial/\partial x + \alpha_2\partial/\partial y + \alpha_3\partial/\partial z) + \beta m]\psi \\ &= (-\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{\nabla} + \beta m)\psi. \end{aligned} \tag{5.19}$$

What are the $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ and β? To solve the requirement $E = +(\boldsymbol{p}^2 + m^2)^{1/2}$, Dirac demanded that his wave function ψ satisfies a KG-type condition (5.15). Then one can show that

---

[167] It means that in the massless case, the left-chiral and right-chiral eigenstates of the chirality operator are also helicity eigenstates, with the same eigenvalues.

[168] For further details, please refer to Section 5.8.

[169] A story told of Dirac, who was known for his eccentricity, is that when he first met Richard Feynman in 1946, he said after a long silence *"I have an equation. Do you have one too?"* Feynman didn't. Dirac walked away after a silence.

Within a few years, Feynman's power as an analyst and intuitionist made him, in the eyes of many, the finest theoretician in America. Wigner agreed with judgement: "*Feynman is a second Dirac, only this time human.*" ([F1]).

The Dirac equation (5.20) appears on the floor of Westminster Abbey on the plaque commemorating Paul Dirac's life, which was unveiled on 13 November 1995.

[170] A conversation in 1928
Bohr : *What are you working on?*
Dirac : *I am trying to get a relativistic theory of the electron.*
Bohr : *But Klein has already solved that problem.*
Dirac (silently) disagreed.

the $\boldsymbol{\alpha}$ and β cannot be ordinary, commuting quantities. Instead they must satisfy the following anticommutation relations ([171])

$$\alpha_i\beta + \beta\alpha_i = 0 \text{ for } i = 1, 2, 3 \text{ and } \alpha_i\alpha_j + \alpha_j\alpha_i = 0 \text{ for } i, j = 1, 2, 3, i \neq j.$$

In addition, it is required that $\alpha_i^2 = \beta^2 = 1$. Now let us multiply the Dirac equation (5.19) with β and due to $\beta^2 = 1$ we get

$$[\mathrm{i}(\beta\partial/\partial t + \beta\boldsymbol{\alpha}\cdot\boldsymbol{\nabla}) - m]\psi = 0.$$

Let us denote $\gamma_0 := \beta$, $\gamma_i := \beta\alpha_i$, $(i = 1, 2, 3)$. Then we can write the Dirac equation in the form

$$(\mathrm{i}\gamma_\mu\nabla^\mu - m)\psi = 0, \text{ ([172])} \tag{5.20}$$

where the coefficients γ_μ satisfy the following relations

$$(\gamma_0)^2 = 1, (\gamma_i)^2 = -1, \ \gamma_\mu\gamma_\nu + \gamma_\nu\gamma_\mu = 0 \tag{5.21}$$

for $\mu \neq \nu$ with $\mu, \nu = 0, 1, 2, 3$ and $i = 1, 2, 3$. ([A1])

Dirac, who had just then been intensely involved with working out the foundations of Heisenberg's matrix mechanics, immediately understood that these conditions could be met if γ_0, γ_1, γ_2 and γ_3 are matrices, acting on a wave function which had several components arranged as a column vector ([A1]). It turns out that the smallest possible dimension of the matrices for which the Dirac conditions can be satisfied is 4×4. One conventional choice of the γ's is

$$\gamma_0 = \begin{bmatrix} \mathrm{I}_2 & 0_2 \\ 0_2 & -\mathrm{I}_2 \end{bmatrix} \qquad \gamma_i = \begin{bmatrix} 0_2 & \sigma_i \\ -\sigma_i & 0_2 \end{bmatrix}, \tag{5.22}$$

where we have written these 4×4 matrices in 2×2 'block diagonal' form, I_2 is the 2×2 identity matrix, 0_2 is the 2×2 null matrix, and σ_i are the so-called 2×2 *Pauli spin matrices*. The Pauli matrices are defined by ([A1])

$$\sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \sigma_2 = \begin{bmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{bmatrix} \quad \sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \tag{5.23}$$

Let us write the gamma matrices ([173]) in the Dirac representation explicitly ([174]):

$$\gamma_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \quad \gamma_1 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}$$

(5.24)

$$\gamma_2 = \begin{bmatrix} 0 & 0 & 0 & -\mathrm{i} \\ 0 & 0 & \mathrm{i} & 0 \\ 0 & \mathrm{i} & 0 & 0 \\ -\mathrm{i} & 0 & 0 & 0 \end{bmatrix} \quad \gamma_3 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

---

[171] These conditions are also required in order to get all the cross-terms such as $\partial x\partial y$ to vanish which in turn is necessary to get a differential equation in first order.

[172] Recall that $\gamma_\mu\nabla^\mu := \sum_{\mu=0}^{3}\gamma_\mu\nabla^\mu = \gamma_0\partial/\partial t + \gamma_1\partial/\partial x + \gamma_2\partial/\partial y + \gamma_3\partial/\partial z$.

[173] Pauli matrices and Dirac matrices are representations of Clifford algebra. A Clifford algebra combines and generalizes the scalar product and the vector product. Pauli matrices define two-dimensional representation of the Clifford algebra of Euclidean space while the Dirac matrices define four-dimensional representation of the Clifford algebra of Minkowski space ([R4]). Clifford (1878) introduced his 'geometric algebras' as a generalization of Grassmann algebras, complex numbers, and quaternions.
William Kingdon Clifford (1845 – 1879) was an English mathematician and philosopher. Hermann Günther Grassmann (1809 – 1877) was a German polymath, known in his day as a linguist and now also as a mathematician.

[174] Please note that despite the μ subscript, the 4-tuple $(\gamma_0, \gamma_1, \gamma_2, \gamma_3)$ is not a four vector. Gamma matrices are constant matrices that remain invariant under a Lorentz transformation.

The conditions (5.21) are usually written in the form

$$\{\gamma_\mu, \gamma_\nu\} = 2\eta^{\mu\nu} I_4, \tag{5.25}$$

where $\mu, \nu = 0, 1, 2, 3$, and

$$\{A, B\} := AB + BA \tag{5.26}$$

is the *anticommutator* of two matrices, $\eta^{\mu\nu}$ is the (μ, ν)-element of the Minkowski metric (see Section 3.1) and I_4 is the 4×4 identity matrix.

Another common choice of the γ's is the Weyl or *chiral representation*, in which γ_i $(i = 1, 2, 3)$ remains the same but γ_0 is different

$$\gamma_0 = \begin{bmatrix} 0_2 & I_2 \\ I_2 & 0_2 \end{bmatrix}.$$

It should be noted that physical results do not depend on the particular 4×4 representation – everything is in the commutation relations. Furthermore, the fact that the Dirac matrices are 4-dimensional is independent of the fact that spacetime is 4-dimensional. The Dirac matrices act on spin space, rather than on spacetime.

Since the Dirac equation involves 4×4 matrices, it is clear that we must interpret the Dirac wave function ψ as a four-component column vector – the so-called *Dirac spinor* (or 4-*spinor* or, sometimes, *bispinor*):

$$\psi = \begin{bmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \\ \psi_4 \end{bmatrix}. \tag{5.27}$$

The four components of ψ can be interpreted as a description of both a spin-½ particle and its antiparticle. The upper two components are associated with the particle, while the lower components are linked to the antiparticle. Each pair of components (upper and lower) can be further decomposed into spin-up and spin-down states.

However, despite of four components, ψ is not a 4-vector since it transforms in a special way under Lorentz transformations. Notice that the Dirac equation (5.20) is simply four coupled differential equations, describing a wave function ψ with four components.

Indeed, we have

$$\gamma_\mu \nabla^\mu = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \partial/\partial t + \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix} \partial/\partial x + \begin{bmatrix} 0 & 0 & 0 & -\mathrm{i} \\ 0 & 0 & \mathrm{i} & 0 \\ 0 & \mathrm{i} & 0 & 0 \\ -\mathrm{i} & 0 & 0 & 0 \end{bmatrix} \partial/\partial y +$$

$$+ \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \partial/\partial z = \begin{bmatrix} \frac{\partial}{\partial \mathrm{t}} & 0 & \frac{\partial}{\partial \mathrm{z}} & \frac{\partial}{\partial \mathrm{x}} - \mathrm{i}\frac{\partial}{\partial \mathrm{y}} \\ 0 & \frac{\partial}{\partial \mathrm{t}} & \frac{\partial}{\partial \mathrm{x}} + \mathrm{i}\frac{\partial}{\partial \mathrm{y}} & -\frac{\partial}{\partial \mathrm{z}} \\ -\frac{\partial}{\partial \mathrm{z}} & -\frac{\partial}{\partial \mathrm{x}} + \mathrm{i}\frac{\partial}{\partial \mathrm{y}} & -\frac{\partial}{\partial \mathrm{t}} & 0 \\ -\frac{\partial}{\partial \mathrm{x}} - \mathrm{i}\frac{\partial}{\partial \mathrm{y}} & \frac{\partial}{\partial \mathrm{z}} & 0 & -\frac{\partial}{\partial \mathrm{t}} \end{bmatrix}.$$

Now we can write out the Dirac equation $(\mathrm{i}\gamma_\mu \nabla^\mu - m)\psi = (\mathrm{i}\gamma_\mu \nabla^\mu - m I_4)\psi = 0_4$ in full

$$\left(\begin{bmatrix} \mathrm{i}\frac{\partial}{\partial t} & 0 & \mathrm{i}\frac{\partial}{\partial z} & \mathrm{i}\frac{\partial}{\partial x}+\frac{\partial}{\partial y} \\ 0 & \mathrm{i}\frac{\partial}{\partial t} & \mathrm{i}\frac{\partial}{\partial x}-\frac{\partial}{\partial y} & -\mathrm{i}\frac{\partial}{\partial z} \\ -\mathrm{i}\frac{\partial}{\partial z} & -\mathrm{i}\frac{\partial}{\partial x}-\frac{\partial}{\partial y} & -\mathrm{i}\frac{\partial}{\partial t} & 0 \\ -\mathrm{i}\frac{\partial}{\partial x}+\frac{\partial}{\partial y} & \mathrm{i}\frac{\partial}{\partial z} & 0 & -\mathrm{i}\frac{\partial}{\partial t} \end{bmatrix} - m\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\right)\begin{bmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \\ \psi_4 \end{bmatrix} =$$

$$\begin{bmatrix} \mathrm{i}\frac{\partial}{\partial t}-m & 0 & \mathrm{i}\frac{\partial}{\partial z} & \mathrm{i}\frac{\partial}{\partial x}+\frac{\partial}{\partial y} \\ 0 & \mathrm{i}\frac{\partial}{\partial t}-m & \mathrm{i}\frac{\partial}{\partial x}-\frac{\partial}{\partial y} & -\mathrm{i}\frac{\partial}{\partial z} \\ -\mathrm{i}\frac{\partial}{\partial z} & -\mathrm{i}\frac{\partial}{\partial x}-\frac{\partial}{\partial y} & -\mathrm{i}\frac{\partial}{\partial t}-m & 0 \\ -\mathrm{i}\frac{\partial}{\partial x}+\frac{\partial}{\partial y} & \mathrm{i}\frac{\partial}{\partial z} & 0 & -\mathrm{i}\frac{\partial}{\partial t}-m \end{bmatrix}\begin{bmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \\ \psi_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

This amounts to the following four equations:

$$\left(\mathrm{i}\frac{\partial}{\partial t}-m\right)\psi_1+\mathrm{i}\frac{\partial}{\partial z}\psi_3+\left(\mathrm{i}\frac{\partial}{\partial x}+\frac{\partial}{\partial y}\right)\psi_4=0, \qquad \left(\mathrm{i}\frac{\partial}{\partial t}-m\right)\psi_2+\left(\mathrm{i}\frac{\partial}{\partial x}-\frac{\partial}{\partial y}\right)\psi_3-\mathrm{i}\frac{\partial}{\partial z}\psi_4=0,$$

$$-\mathrm{i}\frac{\partial}{\partial z}\psi_1+\left(-\mathrm{i}\frac{\partial}{\partial x}-\frac{\partial}{\partial y}\right)\psi_2+\left(-\mathrm{i}\frac{\partial}{\partial t}-m\right)\psi_3=0, \left(-\mathrm{i}\frac{\partial}{\partial x}+\frac{\partial}{\partial y}\right)\psi_1+\mathrm{i}\frac{\partial}{\partial z}\psi_2+\left(-\mathrm{i}\frac{\partial}{\partial t}-m\right)\psi_4=0.$$

A fundamental spinor, often called *Weyl spinor*, has two components (cf. Section 5.7). Thus they see coordinate transformations as 2×2 matrices. There are two kinds of fundamental spinor ψ_L and ψ_R both of which are two-component objects but behave slightly different under coordinate transformations ([175]). They change into one another under a *parity* transformation $(x, y, z) \to (-x, -y, -z)$. Consequently, just one of the Weyl spinors is not sufficient to guarantee parity preservation − we need both of them working together. This is why it is conventional (and convenient) to split the Dirac spinor (5.27) into two fundamental spinors ψ_L and ψ_R ([176]):

$$\psi = \begin{bmatrix} \psi_L \\ \psi_R \end{bmatrix}, \tag{5.27'}$$

where ψ_L and ψ_R are, themselves, two-component column matrices.

Dirac spinors are convenient if we want to make sure that our theory remains valid no matter how we mirror our coordinate axis. Another reason is that one always needs to use a left-chiral spinor and a right-chiral spinor to describe a physical particle like an electron (or a quark). The Dirac spinor applies to massive fermions, all of which are known to obey parity conservation ([177]).

Let us provide a more precise definition of the terms *left-chiral* and *right-chiral* as they relate to wave functions. In order to achieve this, the chirality operator is defined as follows:

---

[175] They are so-called *chirality spinors:* ψ_L is known as *left-chiral spinor* and ψ_R as *right-chiral spinor*. The meaning of the adjective 'chiral' stems from a physical property known as *chirality* (see Section 5.7). In the mathematical language the chirality is a label associated with a representation of the Lorentz group ([S2]).
These spinors are usually called *left-handed* and *right-handed*, respectively. However, this can be misleading because these terms are also used to describe a concept called helicity, which is in general not the same as chirality.

[176] It is convenient to introduce this notation for exactly the same reasons that we introduced 4-vectors. Time and space coordinates of 4-vectors are mixed under Lorentz transformations. Similarly, left-chiral and right-chiral spinors are mixed under coordinate transformations. However, for 4-vectors the mixing happens under boost. In contrast, the mixing of left-chiral and right-chiral spinors happens when we mirror a system. In general, transformations that mirror coordinate axes are known as *parity transformations*. The key observation is that if we consider a transformation law for a right-chiral spinor in a mirrored coordinate system, we find exactly the transformation law of a left-chiral spinor ([S2]).

[177] It turns out that parity is not conserved in weak interactions, which involve neutrinos. There are only left-chiral neutrinos (and right-chiral antineutrinos) – see Section 7.4 for further details.

$$\gamma_5 := \mathrm{i}\gamma_0\gamma_1\gamma_2\gamma_3.$$

Although γ_5 uses the letter gamma, it is not one of the gamma matrices ([W11]). In the chiral representation γ_5 is given by

$$\gamma_5 = \begin{bmatrix} -\mathrm{I}_2 & 0_2 \\ 0_2 & \mathrm{I}_2 \end{bmatrix}.$$

In the Dirac representation, the matrix γ_5 differs from its chiral representation counterpart. Since $(\gamma_5)^2 = \mathrm{I}_4$, the eigenvalues of γ_5 are ± 1. A wave function ψ_L is an eigenstate of γ_5 with eigenvalue -1 ([178]). A right-chiral wave function ψ_R is then an eigenstate with eigenvalue $+1$.

The Dirac equation thus predicts the existence of two types of Dirac wave function: left-chiral functions with chirality -1 and right-chiral functions with chirality $+1$. In order to extract the left- and right-chiral parts of a given wave function, one may define projection operators as follows ([L1])

$$P_L := \frac{\mathrm{I}_4 - \gamma_5}{2}, \quad P_R := \frac{\mathrm{I}_4 + \gamma_5}{2}.$$

P_L and P_R project out the left- and right-chiral components of a spinor ψ:

$$P_L\psi = \begin{bmatrix} \psi_\mathrm{L} \\ 0 \end{bmatrix}, \quad P_R\psi = \begin{bmatrix} 0 \\ \psi_\mathrm{R} \end{bmatrix}.$$

The weak $W^\pm$ couplings contain the projection P_L, and thus only interact with left-chiral particles or right-chiral antiparticles (for further details, please see Section 7.4).

As previously stated in Section 5.7, massive fermions exhibit a coupling between left- and right-chiral wave functions, which is dependent on the mass associated with the particles in question. Let us examine this phenomenon in greater detail.

The Dirac equation can be expressed in the form of two distinct equations ([L1])

$$(\mathrm{I}_2\mathrm{i}\frac{\partial}{\partial t} - \boldsymbol{\sigma}\cdot\nabla)\psi_\mathrm{R} = m\psi_\mathrm{L}, \quad (\mathrm{I}_2\mathrm{i}\frac{\partial}{\partial t} + \boldsymbol{\sigma}\cdot\nabla)\psi_\mathrm{L} = m\psi_\mathrm{R}, \tag{5.28}$$

where $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$ and σ_i $(i = 1, 2, 3)$ are three Dirac matrices.

Equations 5.28 demonstrate that massive fermions comprise both left- and right-chiral wave functions, which are coupled by the particle's mass. We can think of massive Dirac particles oscillating back and forth in time between left- and right-chiral at a rate determined by their mass. This can be most readily observed by considering massive Dirac particles at rest, where we have ([L1])

$$\mathrm{I}_2\mathrm{i}\frac{\partial}{\partial t}\psi_\mathrm{R} = m\psi_\mathrm{L} \quad \text{and} \quad \mathrm{I}_2\mathrm{i}\frac{\partial}{\partial t}\psi_\mathrm{L} = m\psi_\mathrm{R}.$$

In the particular case of a massless particle, the equations (5.28) can be reduced to the following:

$$(\mathrm{I}_2\mathrm{i}\frac{\partial}{\partial t} - \boldsymbol{\sigma}\cdot\nabla)\psi_\mathrm{R} = 0_2, \quad (\mathrm{I}_2\mathrm{i}\frac{\partial}{\partial t} + \boldsymbol{\sigma}\cdot\nabla)\psi_\mathrm{L} = 0_2. \tag{5.28'}$$

This implies that a four-component eigenstate ψ can splits into two two-component pieces, ψ_L and ψ_R, which are not mixed up by the equations. ([L1])

The degree of left- and right-chirality superposition in a Dirac bispinor depends on the energy-to-mass ratio. Consequently, chiral oscillations are typically not relevant for the characterisation of relativistic particles. Nevertheless, they are of considerable importance in the description of particles in dynamical regimes where the momentum is comparable to (or smaller than) the mass. Given that ψ_L and ψ_R are eigenstates of the massless fermion, it can be

[178] Please refer to Section 5.9 for details of the notation used.

concluded that a free and massless left-chiral particle cannot undergo a transformation into a right-chiral particle.

It can be demonstrated ([F14], [L1]) that the massless states ψ_L and ψ_R, which are defined as being chirality eigenstates, are also eigenstates of the helicity operator

$$\hat{h} = \frac{1}{2|\boldsymbol{p}|}\begin{bmatrix} \boldsymbol{\sigma}\cdot\nabla & 0_2 \\ 0_2 & \boldsymbol{\sigma}\cdot\nabla \end{bmatrix},$$

where $\boldsymbol{p}$ is the momentum 3-vector. To be more precise, the left- and right-chirality eigenstates are also helicity eigenstates, with the same eigenvalues. We therefore have the following:

	Eigenvalues	
Particles	**Chirality**	**Helicity**
massless fermions	L = −1	negative = −1 (left-handed)
	R = +1	positive = +1 (right-handed)
massles antifermions	L = +1	positive = +1 (right-handed)
	R = −1	negative = −1 (left-handed)

In conclusion, the concepts of chirality and helicity, despite being represented by different operators, convey a similar message. It is important to note, however, that helicity and chirality are equivalent for massless particles (or when the rest mass can be neglected, e.g. at very high velocity v ≈ c), whereas they are generally not the same for particles with mass. It is evident that helicity provides information regarding the alignment of a particle's spin and momentum, namely whether they are parallel or antiparallel. Consequently, helicity depends on the frame of reference employed for its measurement. In the case of massive particles, it is possible to describe a particle with positive helicity and then to apply a boost to a frame where the particle's momentum is reversed, yet its spin remains unaltered. This results in a reversal of the helicity.

The Dirac equation should be covariant under Lorentz transformations – that is, it must have the same form in the two different frames. In the case of the Klein-Gordon (KG) equation, this requirement is taken care of, almost automatically, by the notation. The case of the Dirac equation is more complicated, because (unlike the KG ψ) the wave function has more than one component, corresponding to the fact that it describes a spinor field related to a spin-½ particle. However, using the matrices γ_μ and Dirac spinors one can construct invariants $\psi^\dagger\gamma_0\psi$ and $\psi^\dagger\gamma_0\gamma_\mu\nabla^\mu\psi$, where $\psi^\dagger$ denotes the *Hermitian conjugate* row vector of the column vector (5.27) ([179]), i.e.

$$\psi^\dagger := [\psi_1^* \;\; \psi_2^* \;\; \psi_3^* \;\; \psi_4^*] = (\psi^T)^*.$$

It is conventional to introduce the *adjoint spinor*

$$\overline{\psi} := \psi^\dagger\gamma_0 = [\psi_1^* \;\; \psi_2^* \;\; \psi_3^* \;\; \psi_4^*]\gamma_0 = [\psi_1^* \;\; \psi_2^* \;\; -\psi_3^* \;\; -\psi_4^*]\,.$$

Now we have everything we need to construct a Lorentz invariant Lagrangian density

$$\mathcal{L}_{Dirac}^{free} = \overline{\psi}(i\gamma_\mu\nabla^\mu - m)\psi. \tag{5.29}$$

Putting this Lagrangian density into the Euler-Lagrange equation one gets the Dirac equation (5.20) which is then Lorentz covariant. Notice that the mass term $m\overline{\psi}\psi$ is invariant under U(1) gauge transformations.

Having set up the relativistic spin-½ free-particle wave equations (5.20) we are now in a position to use the machinery developed in Section 5.4 in order to include electromagnetic interactions. All we have to do is make the replacement ([A1])

$$\nabla^\mu \to D^\mu = \nabla^\mu + iqA^\mu$$

[179] The symbol '†', called *dagger*, denotes transposition plus complex conjugation.

for a particle of charge q:

$$(\mathrm{i}\gamma_\mu D^\mu - m)\psi = 0.$$

The Lagrangian density is in this case given by

(5.30) $$\mathcal{L}_{Dirac} = \mathcal{L}_{Dirac}^{free} + \mathcal{L}_{em}^{free} + \mathcal{L}^{int} = \bar{\psi}(\mathrm{i}\gamma_\mu \nabla^\mu - m)\psi - \frac{1}{4\mu_0} F_{\mu\nu}F^{\mu\nu} - q\bar{\psi}\gamma_\mu A^\mu \psi.$$

The second term with the field $F_{\mu\nu}$ describes the free electromagnetic field, in particular free photons. As we know, $F_{\mu\nu}$ is expressed by derivatives of the vector field $A_\mu(\mathrm{x}_\mu)$ – see (4.7). The interaction term $\mathcal{L}^{int} := -q\,\bar{\psi}\gamma_\mu A^\mu \psi$ is referred to as a *fermion-boson coupling*, or, more specifically, a *fermion-photon interaction*.

This is a Lagrangian density, for example, for an electron and a massless spin-1 boson (photon). In quantum electrodynamics QED every electron is thought to be a localized excitation of the electron (spinor) field $\psi(\mathrm{x}_\mu)$, while every photon is considered to be an excitation of the photon (vector) field $A_\mu(\mathrm{x}_\mu)$, which is the quantum field-theoretic counterpart of the classical four-potential. It is a vector field because photons have spin one. These fields interact and the interactions are quantified by the Lagrangian density $\mathcal{L}^{int}$ ([180]). The Lagrangian density of QED is thus $\mathcal{L}_{QED} := \mathcal{L}_{Dirac} = \bar{\psi}(\mathrm{i}\gamma_\mu D^\mu - m)\psi - \frac{1}{4\mu_0} F_{\mu\nu}F^{\mu\nu}$, which essentially determines everything about the theory ([181]). The covariant derivative D_μ encodes the interaction between the two fields A_μ and ψ, and the 'strength' of the interaction is given by the electric charge q ([182]). It is remarkable that this simple Lagrangian can account for a wide range of phenomena, spanning from those observed at the macroscopic scale down to a length scale of approximately 10^{-13} cm.

Now let us see whether the Dirac equation leads to an acceptable probability current ([A1]). Notice that the quantity

(5.31) $$\rho_{Dirac} := \psi^\dagger(\mathrm{t}, \boldsymbol{x})\psi(\mathrm{t}, \boldsymbol{x}) = \sum_{a=1}^{4} |\psi_a|^2$$

is positive-definite. From the Dirac equation and its Hermitian conjugate one can derive a conservation law of the required form:

$$\nabla \cdot \boldsymbol{j}_{Dirac} = -\partial\rho_{Dirac}/\partial t.$$

The probability current density is

$$\boldsymbol{j}_{Dirac} := \psi^\dagger \boldsymbol{\alpha}\psi$$

representing a 3-vector with components $(\psi^\dagger\alpha_1\psi,\ \psi^\dagger\alpha_2\psi,\ \psi^\dagger\alpha_3\psi) = (\bar{\psi}\gamma_1\psi,\ \bar{\psi}\gamma_2\psi,\ \bar{\psi}\gamma_3\psi)$, where $\boldsymbol{\alpha} = (\alpha_1,\ \alpha_2,\ \alpha_3)$ is the quantity from (5.19). Thus the problem with the negative probability density is solved, because ρ_{Dirac} is positive-definite. ([A1]). The four-vector density current can be written as

(5.31') $$(\mathrm{j}_{Dirac})_\mu := (\rho_{Dirac}, \boldsymbol{j}_{Dirac}) = \bar{\psi}\gamma_\mu\psi.$$

---

[180] More precisely, the interaction term of QED contains three fields (particles): electron and positron $\psi(\mathrm{x}_\mu)$, which is the bispinor Dirac field, and photon $A_\mu(\mathrm{x}_\mu)$. In more general terms, the electrodynamics of a spinor field ψ representing a charged fermion is obtained coupling A_μ to the current $\bar{\psi}\gamma_\mu\psi$. ([M1])

[181] That is, $\mathcal{L}_{QED}$ consists of three parts: the kinetic energy of the electromagnetic field (free photons) $\mathcal{L}_{em}^{free}$, the kinetic energy of the electron field (free electrons) $\mathcal{L}_{em}^{free}$, and the potential energy of their coupling $\mathcal{L}^{int}$.
Or, as Feynman once described it in plain English: *a photon goes from place to place; an electron goes from place to place; an electron emits or absorbs a photon* ([T5]).

[182] For a single electron $q = -1e$, where e is the elementary charge, while e.g. for an up-quark $q = +2e/3$. In case that the field ψ represents an uncharged particle, we have $q = 0$ in (5.30). Then the Lagrangian (5.30) falls into two pieces that have nothing to do with each other. The first term describes the free Dirac field for a particle with mass m. The corresponding field equation is just the Dirac equation (5.20). ([E1])

It turns out, however, that the Dirac equation leads to the same problem as in the case of the Klein-Gordon equation in that for a given value of $\boldsymbol{p}$, two values of E are allowed: $E = \pm(\boldsymbol{p}^2 + m^2)^{1/2}$ i.e. positive and negative energy solutions are still admitted.

Dirac's idea to resolve the dilemma with the negative energy solutions was to introduce a concept where all the states with negative energy are filled with electrons. The completely filled negative-energy states is called the *Dirac sea*, which is invisible by definition. The Pauli exclusion principle then forbids any positive energy electrons from falling into these lower energy levels. If one negative energy electron is absent from the 'Dirac sea', we have a 'hole' relative to the normal vacuum. An electron with positive energy may fall into a hole, which would be observed as a mutual elimination and annihilation of both particles. Conversely, if an unobservable negative energy electron gains sufficient energy to jump to a positive state, a pair of an observable electron and a hole is created. Accordingly, Dirac's theory postulated the existence of a process whereby material particles could be created and annihilated simultaneously. ([K3])

Dirac was inclined to associate this particle with the proton. Of course, the proton mass is approximately 1836 times greater than that of the electron. Nevertheless, Dirac maintained the possibility of deriving the additional mass from the interaction between the hole and the sea, given that the hole moves in the medium of negative energy electrons. ([K3])

The challenges associated with the hole theory, or more specifically, the proposition that the hole is identical to the proton, were becoming more pronounced. The observed mass difference could not be reconciled with the theoretical predictions. In fact, there were compelling arguments in favour of mass equality. Hermann Weyl investigated the mathematical transformations of the theory and reached the conclusion that the hole mass had to be precisely equal to the electron mass. He postulated the existence of a positively charged electron, although he acknowledged that it had not been observed in nature. ([K3])

In May 1931, Dirac submitted a paper in which he discussed two unknown particles. In regard to the difficulties that arose in his hole theory with the proton mass, he abandoned the hypothesis that holes were protons, proposing instead that the theory necessitates the existence of a light positively charged particle, the *anti-electron*. ([K3])

The first announcement of a new positively charged light particle was made by Carl Anderson ([183]) in September 1932. He termed the particle a *positron*. The view that the new particle is nothing other than the anti-electron predicted by Dirac was rapidly accepted as a valid hypothesis. ([K3])

Despite of the experimental evidence for the positron, Dirac's hole theory encountered resistance. In regard to this discovery, Bohr asserted that even if it were proven to be accurate, it would not be consistent with Dirac's theory of holes. In November 1933, Fock published a paper in which he presented a comprehensive and symmetrical analysis of free electrons and positrons, avoiding the use of negative-energy particles. This approach was first proposed by Heisenberg in 1931. This formulation precludes the existence of negative-energy states, which are identified as positive-energy positrons. Therefore, the infinite sea of negative-energy electrons is superfluous. ([V1])

Since Dirac's solution was based on the Pauli exclusion principle, so a few words about it are in order here. In the 1920s it became obvious that the Niels Bohr's model of the atom, proposed in 1913, was inadequate to explain the electron shell structure of an atom. Bohr suggested that electrons could occupy only certain quantized orbitals (designated as shells), but there seemed to be no reason why all the electrons in an atom did not simply crowd into the one lowest energy state. There was no convincing explanation of the structure of the periodic table.

Wolfgang Pauli in 1924 suggested that the pattern in the population of atomic energy levels by electrons could be understood adding a fourth quantum number to the three that were then used to describe an electron's quantum state. The first three quantum numbers made sense

---

[183] Carl Anderson (1905 – 1991) was an American physicist. He received 1936 Nobel Prize in Physics for his discovery of the positron.

physically, since they related to the electron's motion around the nucleus. Pauli called his new quantum property of the electron a *"two-valuedness not describable classically"*. Then in January 1925, he announced the exclusion principle, stating that no two electrons in an atom can occupy a state with the same values for the four quantum numbers. Each electron had to be in its own unique state. Other possibilities are excluded. ([A2])

This new quantum property was understood a little later as due to the spin of the electron (see Section 5.7). The exclusion principle answered a question that had remained obscure in the old atomic theory of Bohr and Sommerfeld: why do not all the electrons in atoms fall down into the shell of lowest energy? Subsequently Pauli's exclusion principle was incorporated into statistical mechanics by Fermi and Dirac ([184]) and for this reason particles obeying the exclusion principle are generally called 'fermions' (cf. Chapter 2) ([W1]). On the basis of relativistic quantum field theory only, Fierz ([185]) (1939) and Pauli (1940) proved that all particles are either bosons or fermions ([E1]).

Applying the exclusion principle proved possible to obtain sensible results from the Dirac equation and its negative energy solutions. It is clear, however, that the theory is no longer really a 'single-particle' theory. For example, if we excite one negative energy electron to a positive energy state, we have in the final state a positive energy electron plus a positive energy positron 'hole' in the vacuum: this corresponds physically to the process of e^+e^- pair creation. Thus this way of dealing with the negative energy problem for fermions leads us directly to the need for a quantum field theory ([A1]). In QED, ψ becomes an operator capable of creating or annihilation of e^+e^- pairs.

The *positron* e^+ or *anti-electron* is the antiparticle of the electron e^-. The positron has an electric charge of $+1e$, a spin of $\frac{1}{2}$ (the same as the electron), and has the same mass as the electron. Fermions (like electron) that are not their own antiparticles are referred to as *Dirac fermions*. The term is sometimes used in opposition to a *Majorana fermion*, which is a fermion that is its own antiparticle. Such fermions were hypothesised by Ettore Majorana ([186]) in 1937. Except for the neutrino, all of the Standard Model fermions are known to behave as Dirac fermions at low energy (after electroweak symmetry breaking), and none are Majorana fermions. The nature of the neutrinos is not settled – they may be either Dirac or Majorana fermions ([W11]). ([187])

5.9. An interlude on the quantum formalism ([188])

This section provides an overview of the quantum formalism. It is a framework that permits the mathematical description of quantum systems. The initial fundamental components are new entities, designated as 'ket' and 'bra' used to describe quantum states of the system under

---

[184] In late 1925, Jordan submitted a manuscript to Born with a request for publication in the *Zeitschrift für Physik*, which was then under Born's editorship. Born then went on a long trip to the United States and did not remember the paper he had put in his suitcase. In the meantime, Fermi-Dirac statistics had been discovered independently by Fermi and Dirac. But Jordan was the first. It is worth noting that Jordan referred to his discovery as Pauli statistics ([E1]). Ernst Pascual Jordan (1902–1980) was a German theoretical and mathematical physicist who made significant contributions to quantum mechanics and quantum field theory.

[185] Markus Eduard Fierz (1912–2006) was a Swiss physicist.

[186] Ettore Majorana (1906 – probably died after 1959) was an Italian theoretical physicist who worked on neutrino masses. He was very young when he joined Enrico Fermi's team in Rome as one of the 'Via Panisperna boys', who took their name from the street address of their laboratory. On 25 March 1938, he disappeared under mysterious circumstances while going by ship from Palermo to Naples. Despite several investigations, his body was not found and his fate is still uncertain. Italian philosopher Giorgio Agamben published in 2016 a book that examines the case of Majorana's disappearance ([W11]).

[187] Majorana suggested that neutral spin-$\frac{1}{2}$ particles can be described by a real-valued wave equation (the *Majorana equation*), and would therefore be identical to their antiparticle since the wave functions of particle and antiparticle are related by complex conjugation.

[188] The content of this section is to a great extent borrowed from the book [S16] that I highly recommend to the reader for further details. This book is the ultimate practical introduction to quantum mechanics.

examination. Furthermore, it is necessary to define operators that yield, for instance, the momentum of the system when they are applied to a given ket. As will be demonstrated subsequently, the measured values of physical quantities (e.g. momentum or position) equate to the eigenvalues of the corresponding operators. ([S2a])

Quantum mechanics represents such a radical departure from classical physics that the very notion of the state of a particle must be changed. The space of states of a quantum system is not merely a collection of potential outcomes; rather, it can be conceptualised as a vector space. A mathematical vector space is an abstract construction that may or may not have any relation to the physical space we perceive. The number of dimensions may vary from one to infinity, and the components may be real numbers, complex numbers, or even more general entities.

The vector spaces used to define quantum mechanical states are called *Hilbert spaces* ([189]). When you come across the term Hilbert space in quantum mechanics, it refers to the space of states. A Hilbert space may have either a finite or an infinite number of dimensions.

The simplest example of a Hilbert space is $\mathbb{C}^n := \mathbb{C} \times \mathbb{C} \times \ldots \times \mathbb{C}$ (n-times), where $\mathbb{C}$ denotes the set of complex numbers, with the usual inner product (also called *Hermitian product*)

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle := \boldsymbol{u}^{\dagger} \boldsymbol{v} = (\boldsymbol{u}^{\mathrm{T}})^{*} \boldsymbol{v} = \sum_{j=1}^{n} u_j^{*} v_j,$$

where $\boldsymbol{u} = (u_1, \ldots, u_n)$, $\boldsymbol{v} = (v_1, \ldots, v_n) \in \mathbb{C}^n$ and u_j^{*} denotes the complex conjugate of u_j. Notice that $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \langle \boldsymbol{v}, \boldsymbol{u} \rangle^{*}$ since $(z + w)^{*} = z^{*} + w^{*}$ and $(z^{*})^{*} = \mathrm{z}$ for any complex numbers *z* and *w*.

The space $\mathbb{C}^n$ is a finite dimensional Hilbert space because it possesses a finite *basis* which is for example $(\boldsymbol{e}_j)_{j=1}^{n}$ where

$$\begin{aligned} \boldsymbol{e}_1 &= (1, 0, \ldots, 0, 0, 0, \ldots, 0), \\ \boldsymbol{e}_2 &= (0, 1, \ldots, 0, 0, 0, \ldots, 0), \\ &\ldots \\ \boldsymbol{e}_k &= (0, 0, \ldots, 0, 1, 0, \ldots, 0), \\ &\ldots \\ \boldsymbol{e}_n &= (0, 0, \ldots, 0, 0, 0, \ldots, 1). \end{aligned}$$

Every vector $\boldsymbol{u} = (u_1, \ldots, u_n) \in \mathbb{C}^n$ can be written as a linear combination of the elements of the basis

$$\boldsymbol{u} = \sum_{j=1}^{n} u_j \, \boldsymbol{e}_j. \tag{5.32}$$

Notice that

$$\langle \boldsymbol{e}_j, \boldsymbol{u} \rangle = u_j, \tag{5.33}$$

for $j = 1, 2, \ldots,$ n ([190]). This is analogous to how we expand an arbitrary 3-vector $\boldsymbol{u} = (u_x, u_y, u_z)$ in the terms of basis vectors

---

[189] Hilbert spaces are named after German mathematician David Hilbert (1862 – 1943), who studied them in the context of integral equations. A Hilbert space $\mathcal{H}$ is a (complex) vector space that is equipped with an inner product. This is a map $\langle\,,\rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{C}$ such that for *u, v, w* $\in \mathcal{H}$ and $\lambda \in \mathbb{C}$, (1) $\langle u, v \rangle = \langle v, u \rangle^{*}$; (2) $\langle u, \lambda v \rangle = \lambda \langle u, v \rangle$; (3) $\langle u, v + w \rangle = \langle u, v \rangle + \langle u, w \rangle$; (4) $\langle u, u \rangle \geq 0$ and $\langle u, u \rangle = 0$ if and only if u = 0. Remainder: $\langle v, u \rangle^{*}$ denotes the complex conjugate of $\langle v, u \rangle$. Note that (1) and (2) imply $\langle \lambda u, v \rangle = \lambda^{*} \langle u, v \rangle$ so that $\langle\,,\rangle$ is antilinear in its first (leftmost) argument.

In the maths literature, the inner product is often taken to be linear in the left entry and antilinear in the right. We will follow the QM literature, which always uses the opposite convention to the maths literature.

Using this inner product we define the *norm* of a vector $u \in \mathcal{H}$ to be $||u|| = \sqrt{\langle u, u \rangle}$ which is assumed to be complete.

Thus every Hilbert space is a *Banach space*. Moreover, $\mathcal{H}$ is in QM assumed to be *separable*, i.e. it has a countable dense subset. There are several important reasons for this assumption: Firstly, the separability of the Hilbert space ensures that the set of possible measurement outcomes is countable, which aligns with the physical reality of discrete measurement results. Secondly, the separability condition guarantees that the probability measures used in quantum mechanics are well-defined and can be normalized, ensuring a consistent probabilistic interpretation of quantum states.

[190] The basis $(\boldsymbol{e}_j)_{j=1}^{n}$ is *orthonormal*. It means that it has the properties: $\langle \boldsymbol{e}_j, \boldsymbol{e}_j \rangle = 1$ and $\langle \boldsymbol{e}_j, \boldsymbol{e}_k \rangle = 0$ for $j \neq k$. Mathematically, basis vectors are not required to be orthonormal. However, in quantum mechanics they generally are.

$$\boldsymbol{e}_x = (1, 0, 0), \boldsymbol{e}_y = (0, 1, 0), \boldsymbol{e}_z = (0, 0, 1).$$

In general, the precise nature of the Hilbert space depends on the specific quantum system under consideration. To illustrate, the Hilbert space for the spin of a single particle is the space C^2 of two-dimensional complex vectors (for further details, see Example 5.2 below). An additional example is the space of complex *square-integrable* ([191]) functions $L^2(\mathbb{C})$, which is used for the description of a particle's position and momentum. The question thus arises as to which Hilbert space is the correct one to use in order to describe a particular system under consideration. Ultimately, this question can only be determined by carrying out an experiment.

In quantum mechanics everything we could want to know about a physical system is encoded in a vector ([192]) in a Hilbert space $\mathcal{H}$. From now on, in this section we will use a notation for Hilbert spaces that was introduced by Dirac and is standard throughout the theoretical physics literature. Dirac denotes an element of $\mathcal{H}$ as $|\Psi\rangle$, where the symbol $|\ \rangle$ is known as a *ket* ([D2]). An element of the dual space, that is essentially the complex conjugate vector space, is written $\langle\Psi|$ and the symbol $\langle\ |$ is called a *bra*. It is often convenient to think of $|\Psi\rangle$ as represented by a column vector ([193])

$$|\Psi\rangle = \begin{bmatrix} \Psi_1 \\ \Psi_1 \\ \dots \\ \Psi_n \end{bmatrix},$$

and $\langle\Psi|$ by a row vector with the components being complex conjugates

$$\langle\Psi| = [\Psi_1^* \quad \Psi_2^* \quad \dots \quad \Psi_n^*].$$

Thus the relation between the ket $|\Psi\rangle$ and the bra $\langle\Psi|$ is

$$\langle\Psi| := |\Psi\rangle^\dagger = (|\Psi\rangle^{\mathrm{T}})^* \ ([194])$$

The inner product between two states $|\Psi\rangle$ and $|\Phi\rangle$ is then written $\langle\Psi|\Phi\rangle$ forming a *bra-ket* or *bracket*. Of course the components Φ_j and Ψ_j depend on the basis of $\mathcal{H}$. The inner product $\langle\Psi|\Phi\rangle$ however, is independent of the choice of basis.

In quantum mechanics, two state vectors $|\Psi\rangle$ and $\lambda\,|\Psi\rangle$, where λ is any nonzero complex number, have exactly the same physical significance ([195]). For this reason, it is sometimes helpful to say that the physical state corresponds not to a particular vector in the Hilbert space, but to the ray, or one-dimensional subspace, defined by the set of all the complex multiples of a particular vector. Consequently, one can always choose λ (recall that $|\Psi\rangle$ is not the zero vector) in such a way that the $|\Psi\rangle$ corresponding to a particular physical situation is normalised, $\langle\Psi|\Psi\rangle = 1$ or $||\Psi|| = 1$, where the norm $||\Psi||$ of a state $|\Psi\rangle$ is the square root of $\langle\Psi|\Psi\rangle$, i.e.

$$||\Psi||^2 = \langle\Psi|\Psi\rangle = \sum_{j=1}^{n} \Psi_j^* \Psi_j.$$

Two states $|\Psi\rangle$ and $|\Phi\rangle$ are *distinguishable* if they are orthogonal, i.e. if their inner product is zero: $\langle\Psi|\Phi\rangle = 0$. This is the analogue of saying that two 3-vectors are orthogonal if their dot product is zero.

---

[191] In mathematics, a square-integrable function, also called a quadratically integrable function, is a real- or complex-valued measurable function for which the integral of the square of the absolute value is finite.

[192] This vector is not the zero vector, since the zero vector never represents any physical situation.

[193] If $\mathcal{H}$ is finite-dimensional.

[194] More exactly, the bra vector is a vector in the dual space $\mathcal{H}^*$, i.e. in the vector space of all continuous linear functionals $f : \mathcal{H} \to \mathbb{C}$. We can however identify $\mathcal{H}$ with $\mathcal{H}^*$ due to the Riesz representation theorem. This is special about Hilbert spaces among various other infinite dimensional vector spaces, and makes them especially easy to handle.

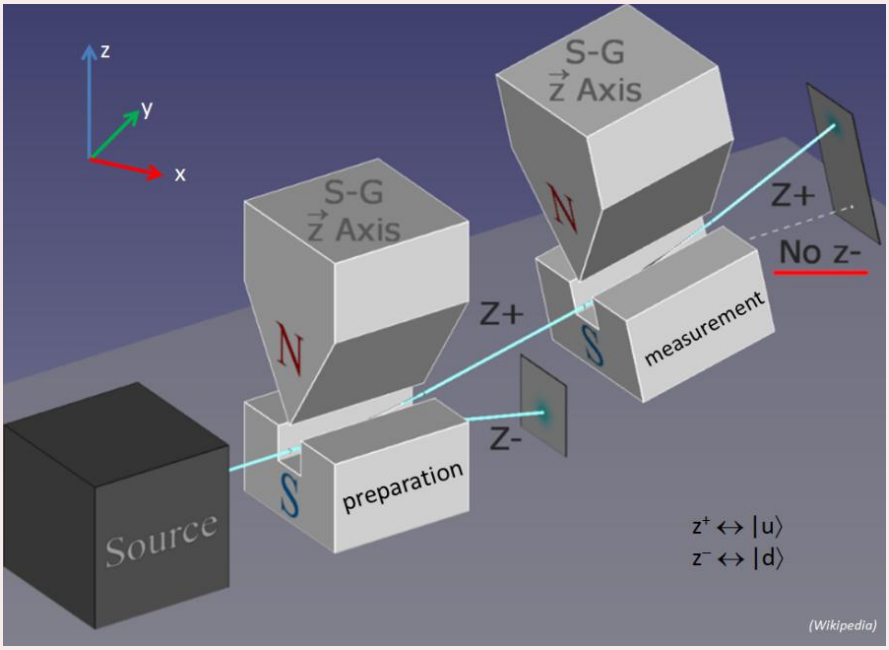
[195] Avoid, however, the following mistake. Just because vectors $|\Psi\rangle$ and $\lambda\,|\Psi\rangle$ have the same physical interpretation does not mean that one can multiply a vector inside some formula by a constant without changing the physics. An overall constant makes no difference, but changing the relative magnitudes or phases of two kets for example in a sum can make a difference.

Although for many purposes it is convenient to use normalised vectors, one should not get the mistaken impression that any vector representing a quantum system must be normalised. There are circumstances in which it is more convenient not to use normalised vectors.

Normalised vectors can always be multiplied by a phase factor, a complex number of the form $e^{i\alpha}$ where α is real, without changing the normalisation or the physical interpretation, so normalisation by itself does not single out any unique vector representing a particular physical state.

Example 5.2. Let us see how to represent spin states of a particle, say an electron, using state vectors. Electron has, besides spin, a magnetic momentum proportional to the spin. So, if we measure the magnetic momentum, we get some idea about the spin. As we know, spin is quantised, and can only take on discrete values. The spin angular momentum of an electron, measured along any particular direction, can only take on the values $\hbar/2$ or $-\hbar/2$. Since we assume $\hbar = 1$, we have two spin values: $+\frac{1}{2}$ and $-\frac{1}{2}$. There are no intermediate values.

Let us begin by labelling the possible spin states along the three coordinate axes. A physical picture is to think of the spin-½ particle as having an angular momentum vector pointing in a random direction in space, but subject to the constraint that a particular component of the angular momentum, say S_z along the z-axis, is described as *spin up* or *spin down*, based on the magnetic momentum pointing up or down, respectively. Let us denote these states by ket vectors $|u\rangle$ and $|d\rangle$. One can prepare an electron with the spin in a certain direction and measure it using a Stern-Gerlach device. Thus, when our detector is oriented along the z-axis and registers $+\frac{1}{2}$, the state $|u\rangle$ has been prepared ([W11]):



So we have a very simple mathematical representation: all possible spin states can be represented in a two-dimensional vector space $\mathbb{C}^2$. Let $|e_1\rangle$ and $|e_2\rangle$ be two basis vectors in $\mathbb{C}^2$ and assume that both of them are normalised (i.e. $\langle e_1, e_1\rangle = \langle e_2, e_2\rangle = 1$) and that they are mutually orthogonal (i.e. $\langle e_1, e_2\rangle = \langle e_2, e_1\rangle = 0$). For example, we can take ([196])

$$|e_1\rangle := \begin{bmatrix} 1 + i0 \\ 0 + i0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad |e_2\rangle := \begin{bmatrix} 0 + i0 \\ 1 + i0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The usual convention is

[196] Notice that these vectors are actually spinors - see Section 5.7.

$$|e_1\rangle = |u\rangle \leftrightarrow S_z = +\frac{1}{2}, \quad |e_2\rangle = |d\rangle \leftrightarrow S_z = -\frac{1}{2}.$$

Then we can express any spin state as a linear superposition of $|u\rangle$ and $|d\rangle$. And each (non-zero) vector in this vector space represents a possible state of the electron spin. We are not discussing other degrees of freedom of the electron, such as its position, momentum, or energy.

Denoting by $|\Psi\rangle$ a generic spin state we can write this linear superposition as an equation

$$|\Psi\rangle = a_u|u\rangle + a_d|d\rangle, \tag{5.34}$$

where a_u and a_d are the components of $|\Psi\rangle$ along the basis directions $|u\rangle$ and $|d\rangle$. Applying (5.33), we can identify the components of $|\Psi\rangle$ as

$$a_u = \langle u, \Psi\rangle \quad \text{and} \quad a_d = \langle d, \Psi\rangle.$$

What is the physical significance of these equations? The vector $|\Psi\rangle$ can represent any state of the spin, prepared in any manner. The components a_u and a_d are complex numbers and by themselves have no experimental meaning, but their magnitudes do. For example, $|a_u|^2 = a_u^* a_u$ means the following: given that the spin has been prepared in the state $|\Psi\rangle$, and that the detector is oriented along the z-axis, the quantity $|a_u|^2$ is the probability that the spin would be measured as $S_z = +\frac{1}{2}$. In other words, it is the probability of the spin being up if measured along the z-axis.

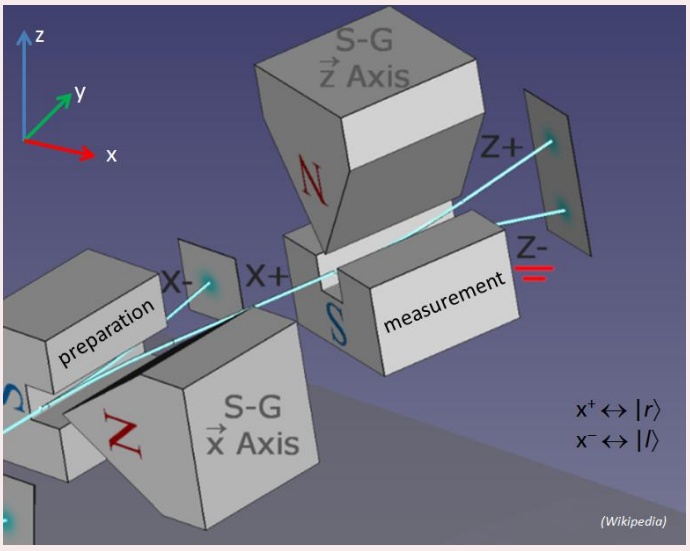
The important point is that before the measurement the vector $|\Psi\rangle$ represents the potential possibilities but not the actual values of our measurements. Since $|u\rangle$ and $|d\rangle$ are mutually orthogonal, we have $\langle u, \mathrm{d}\rangle = \langle d, \mathrm{u}\rangle = 0$. The physical meaning of this is that, if the spin is prepared up, then the probability to detect it down is zero, and vice versa. It implies that two orthogonal states are physically distinct and mutually exclusive. And this idea applies to all quantum systems, not just spin.

The second important point is that the total probability must equal to unity, so we must have

$$|a_u|^2 + |a_d|^2 = a_u^* a_u + a_d^* a_d = 1.$$

This is equivalent to the condition that the vector $|\Psi\rangle$ is normalised: $\langle \Psi, \Psi\rangle = 1$.

Now let us consider the spin component S_x along x-axis. If the apparatus is oriented along the x-axis and registers $-\frac{1}{2}$, the state $|l\rangle$ has been prepared. We will call it *spin left*. The second possible state that can be prepared corresponds to the state $|r\rangle$, which we call *spin right*. According to (5.34) we can represent any spin state as a linear combination of the basis vectors $|u\rangle$ and $|d\rangle$. How can we represent e.g. the vector $|r\rangle$ in this basis? It turns out that if our apparatus initially prepares $|r\rangle$, and is then rotated to measure S_z, there will be equal probabilities for up and down ([W11]):



Thus, $a_u^* a_u = a_d^* a_d = \frac{1}{2}$ and a vector $|r\rangle$ that satisfies this rule is

$$|r\rangle = \frac{1}{\sqrt{2}}|u\rangle + \frac{1}{\sqrt{2}}|d\rangle.$$

One can show that the vector $|l\rangle$ has the following representation in the basis $|u\rangle$ and $|d\rangle$

$$|l\rangle = \frac{1}{\sqrt{2}}|u\rangle - \frac{1}{\sqrt{2}}|d\rangle.$$

Finally, the vectors representing spins oriented *in* and *out* along the y axis are

$$|i\rangle = \frac{1}{\sqrt{2}}|u\rangle + \frac{\mathrm{i}}{\sqrt{2}}|d\rangle, \quad |o\rangle = \frac{1}{\sqrt{2}}|u\rangle - \frac{\mathrm{i}}{\sqrt{2}}|d\rangle.$$

Notice that two of the components in the above equations contain i ($\mathrm{i}^2 = -1$), i.e. they are imaginary. Given our framework for spin states, there is no way around them. The need for complex numbers is a general feature of quantum mechanics.

The last experiment can be interpreted to exhibit the uncertainty principle: since the angular momentum cannot be measured on two perpendicular directions at the same time, the measurement of the angular momentum on one direction destroys the previous determination of the angular momentum in the other direction.

We have seen in the example above how states in quantum mechanics can be mathematically described as vectors in a vector space. The next step in quantum mechanics is to introduce the idea of an *observable*. An observable could also be called a *measurable*. It is a thing that you can measure with a suitable apparatus. For example, measuring the components of a spin, S_x, S_y and S_z. Or we can make measurements of the coordinates of a particle, the energy or momentum. These are examples of observables.

Observables are also associated with a vector space, but they are not state vectors. They are the things you measure and they are mathematically represented by linear operators acting on a vector space of states. The correspondence between operators and observables is subtle, and understanding it requires some effort. ([197])

In order to describe this correspondence, we need to discuss linear operators in little more detail. We have already come upon linear transformations (= operators) in Section 3.5. An operator T acts on a vector, say $|\Psi\rangle$, and gives another vector, say $|\Phi\rangle$: $T|\Psi\rangle = |\Phi\rangle$. We require that T gives a unique output for every vector in the space. Recall that T: $\mathcal{H} \to \mathcal{H}$ is linear if

$$T(a|\Psi\rangle + b|\Phi\rangle) = aT|\Psi\rangle + bT|\Phi\rangle,$$

for any complex numbers a and b. The set of all linear operators is itself a vector space, since a scalar times an operator is an operator, and the sum of two operators is also an operator. The operator $aT + bR$ applied to an element $|\Psi\rangle$ of $\mathcal{H}$ yields the result:

$$(aT + bR)|\Psi\rangle = aT|\Psi\rangle + bR|\Psi\rangle.$$

The product TR of two operators T and R is the operator obtained by first applying R to some ket, and then T to the ket which results from applying R:

$$TR(|\Psi\rangle) = T(R(|\Psi\rangle)).$$

---

[197] The reader may ask why a physical quantity is represented by its associated operator. The initial information the observer has about a quantum system comes from a set of measurements. This is the same as in classical physics. The state of the system represents this information, which can be cast into different mathematical forms. It is usually described in terms of a state vector or a wave function. The wave function has no direct physical meaning. It is just one way of storing information. It stores all the information available to the observer about the system. To make predictions about the outcome of all measurements, at any time, one has to 'do' something to the wave function to extract the information. One has to perform some mathematical operation on it, such a squaring it, multiplying it by a constant, differentiating it, etc. One has to operate on the wave function with some operator. The operator is a specific instruction or set of instructions. Every observable is associated with its own operator ([B8]). We use linear operators because we see quantum effects that exhibit linear superposition of states, and linear operators are the right mathematical objects for dealing with such superposition.

Normally the parentheses are omitted, and one simply writes $TR|\Psi\rangle$. However, it is very important to note that operator multiplication, unlike multiplication of scalars, is not commutative: in general, $TR \neq RT$. In the exceptional case in which $TR = RT$ one says that these two operators *commute with each other*, or (simply) *commute*.

If the Hilbert space under consideration is finite-dimensional, say $\mathcal{H} = \mathbb{C}^n$, then a linear operator $T:\mathcal{H} \to \mathcal{H}$ can be represented by an n×n matrix ([198])

$$T = \begin{bmatrix} t_{11} & t_{12} & \dots & t_{1n} \\ t_{21} & t_{22} & \dots & t_{2n} \\ \dots & \dots & \dots & \dots \\ t_{n1} & t_{n2} & \dots & t_{nn} \end{bmatrix}.$$

Thus in this case we can write $T|\Psi\rangle = |\Phi\rangle$ as

$$\begin{bmatrix} t_{11} & t_{12} & \dots & t_{1n} \\ t_{21} & t_{22} & \dots & t_{2n} \\ \dots & \dots & \dots & \dots \\ t_{n1} & t_{n2} & \dots & t_{nn} \end{bmatrix} \begin{bmatrix} \Psi_1 \\ \Psi_1 \\ \dots \\ \Psi_n \end{bmatrix} = \begin{bmatrix} \Phi_1 \\ \Phi_1 \\ \dots \\ \Phi_n \end{bmatrix}. \tag{5.35}$$

The inner product of some element $|\Phi\rangle$ of $\mathcal{H}$ with the ket $T|\Psi\rangle$ can be written as

$$|\Phi\rangle^\dagger T|\Psi\rangle = \langle\Phi|T|\Psi\rangle$$

where the notation on the right side, the 'sandwich' with the operator between a bra and a ket, is standard Dirac notation. It is often referred to as a 'matrix element', even when no matrix is actually under consideration. One can write $\langle\Phi|T|\Psi\rangle$ as $(\langle\Phi|T)(|\Psi\rangle)$, and think of it as the linear functional or bra vector $\langle\Phi|T$ acting on or evaluated at $|\Psi\rangle$. Of course, $\langle\Phi|T|\Psi\rangle \in \mathbb{C}$.

In this sense, it is natural to think of a linear operator T on $\mathcal{H}$ as inducing a linear map of the dual space $\mathcal{H}^*$ onto itself, which carries $\langle\Phi|$ to $\langle\Phi|T$. This map can also, without risk of confusion, be denoted by T, and while one could write it as $T(\langle\Phi|)$, in Dirac notation $\langle\Phi|T$ is more natural. Sometimes one speaks of 'the operator T acting to the left'. In this way, linear operators can also act on bra-vectors. If T is represented by a matrix, then $\langle\Psi|T$ stands just for multiplying $\langle\Psi|$ by T.

Given an operator $T:\mathcal{H} \to \mathcal{H}$, its *Hermitian* ([199]) *conjugate* $T^\dagger$ is the unique operator such that

$$\left\langle\Phi\middle|T^\dagger\middle|\Psi\right\rangle = \langle\Phi|T|\Psi\rangle^*$$

for any $|\Phi\rangle$ and $|\Psi\rangle$ in $\mathcal{H}$ ([200]).

Quantum mechanical observables are represented by a special kind of linear operators. They are represented by linear operators that are equal to their own Hermitian conjugates. Such operators are called *Hermitian operators*. If a linear operator is represented by a matrix U then it is Hermitian provided $\mathrm{U} = \mathrm{U}^\dagger := (\mathrm{U}^\mathrm{T})^* = (\mathrm{U}^*)^\mathrm{T}$, i.e. if U is equal to its own conjugate transpose. In terms of matrix elements, this can be written as $u_{jk} = u_{kj}^*$.

Why do we use in quantum mechanics only Hermitian operators for representing observables? We will discuss this question in a moment. First we need to introduce the notions of eigenvalues and eigenvectors (= eigenstates).

In general, when a linear operator acts on a vector, it can change its direction and/or its magnitude. For a particular linear operator, there are however certain vectors whose directions do not change. These special vectors are called *eigenvectors* or *eigenstates*. Thus $|\Psi\rangle$ is an eigenvector of T if

---

[198] Equating a linear operator with a matrix, which depends on a particular basis, is sloppy but it should not cause confusion.

[199] After the French mathematician Charles Hermite (1822–1901).

[200] Note that $T^\dagger$: $\mathcal{H}^* \to \mathcal{H}^*$, where $\mathcal{H}^*$ is the dual space of $\mathcal{H}$.

(5.36) $$T|\Psi\rangle = \lambda|\Psi\rangle,$$

where λ is a number, known as the *eigenvalue* associated with $|\Psi\rangle$.

Hermitian operators have the following important properties that build a foundation of quantum mechanics.

- The eigenvectors of a Hermitian operator $T{:}\mathcal{H} \to \mathcal{H}$ form a basis of the space $T(\mathcal{H})$. This means that any vector, the operator can generate, can be expanded as a sum (i.e. linear combination) of its eigenvectors.
- Eigenvalues of T are real.
- If λ_1 and λ_2 are two different eigenvalues of T, then the corresponding eigenvectors are orthogonal.
- Even if the two eigenvalues are equal, the corresponding eigenvectors can be chosen to be orthogonal. This situation, where two different eigenvectors have the same eigenvalue, is called *degeneracy* (we shall consider, for example, energy degeneracy in Section 6.1).

In quantum mechanics the operator corresponding to an observable Q is usually denoted by $\hat{Q}$. When we measure the observable Q, the possible results of a measurement are the eigenvalues of the operator $\hat{Q}$. Since the result of an experiment must be a real number, the eigenvalues of the operator $\hat{Q}$ must also be real. Moreover, the eigenvectors that represent unambiguously distinguishable results must have different eigenvalues, and must also be orthogonal. These conditions are sufficient to prove that $\hat{Q}$ must be Hermitian.

Let us repeat some important facts. First, when an observable is measured, the result is always a real number drawn from a set of possible results. For example, if the energy of an atom is measured, the result will be one of the established energy levels of the atom. In case of the spin (of a fermion), the possible values of any of the spin components are $\pm\frac{1}{2}$. The apparatus never gives any other result. Moreover, the result of a measurement is generally statistically uncertain. However, for any given observable, there are particular states for which the result is absolutely certain. These states correspond to the eigenvectors of the operator representing the observable. For example, if the S-G apparatus is oriented along the z-axis, the state $|u\rangle$ never gives anything but $S_z = +\frac{1}{2}$.

Example 5.3. In this example, we shall look more closely at the form of the operators representing observables which are the spin components S_x, S_y and S_z.

Let the operator $\hat{S}_a$ represents the observable S_a, $a = x, y, z$. We know that the operator $\hat{S}_a$: $\mathbb{C}^2 \to \mathbb{C}^2$ must Hermitian, i.e. it has a representation as Hermitian 2×2 complex matrix. What is this matrix?

Let us begin with the operator $\hat{S}_z$. We know that it has definite, unambiguous values for the states $|u\rangle$ and $|d\rangle$, and that the corresponding measurement values are $+\frac{1}{2}$ and $-\frac{1}{2}$. Hence applying (5.36) we get

(5.37) $$\hat{S}_z|u\rangle = \frac{1}{2}|u\rangle \quad \text{and} \quad \hat{S}_z|d\rangle = -\frac{1}{2}|d\rangle.$$

Moreover, states $|u\rangle$ and $|d\rangle$ are orthogonal to each other, i.e. $\langle u, d\rangle = 0$. Using this condition and replacing in (5.37) $\hat{S}_z$ by a 2×2 matrix, $|u\rangle$ and $|d\rangle$ by the column vectors

$$|u\rangle = \begin{bmatrix}1\\0\end{bmatrix} \quad \text{and} \quad |d\rangle = \begin{bmatrix}0\\1\end{bmatrix},$$

respectively, we infer that

(5.38) $$\hat{S}_z = \frac{1}{2}\begin{bmatrix}1 & 0\\0 & -1\end{bmatrix}.$$

Thus $\hat{S}_z$ as defined in (5.38) is a quantum mechanical operator representing the observable S_z. This operator has two eigenvectors $|u\rangle$ and $|d\rangle$ with associated eigenvalues $+\frac{1}{2}$ and $-\frac{1}{2}$, respectively. We do the same for the other two components of spin, S_x and S_y to obtain

(5.39) $$\hat{S}_x = \frac{1}{2}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \hat{S}_y = \frac{1}{2}\begin{bmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{bmatrix}.$$

We see now that the quantum mechanical spin operators are represented by the Pauli matrices (5.23)

$$\hat{S}_x = \frac{1}{2}\sigma_1, \quad \hat{S}_y = \frac{1}{2}\sigma_2, \quad \hat{S}_z = \frac{1}{2}\sigma_3.$$

In Example 5.2 spin components were measured by orienting our S-G apparatus along any one of the three coordinate axes x, y and z. But of course we can orient the apparatus along any axis and measure the spin component along it. In other words, take any unit 3-vector $\boldsymbol{n}$ and orient the S-G apparatus along $\boldsymbol{n}$. Activating our apparatus would then measure the spin component $S_{\boldsymbol{n}}$ along the axis $\boldsymbol{n}$. What is an operator $\hat{S}_{\boldsymbol{n}}$ that corresponds to this observable? Assuming that $\boldsymbol{n}$ is a unit 3-vector with components n_1, n_2, n_3 one can show that $\hat{S}_{\boldsymbol{n}}$ is a linear combination of the Pauli matrices

(5.40) $$\hat{S}_{\boldsymbol{n}} = \frac{1}{2}\boldsymbol{n}\cdot\boldsymbol{\sigma} = \frac{1}{2}(n_1\sigma_1 + n_2\sigma_2 + n_3\sigma_3).$$

Using (5.40) we can for a given vector $\boldsymbol{n}$ compute eigenvectors and eigenvalues of the operator $\hat{S}_{\boldsymbol{n}}$. Then we will know the possible outcomes of a measurement along the direction of $\boldsymbol{n}$. And we will also be able to calculate probabilities for those outcomes. In other words, we will have a complete picture of spin measurements in three-dimensional space.

It is important to point out that measuring an observable is not the same as operating with the corresponding operator on the state. In particular, it is in general wrong to say that if the state of the system before we do the measurement is $|\Psi\rangle$, then the measurement of Q changes the state to $\hat{Q}|\Psi\rangle$. To see this let us consider, as an example, the prepared spin state

$$|r\rangle = \frac{1}{\sqrt{2}}|u\rangle + \frac{1}{\sqrt{2}}|d\rangle.$$

Acting on this state-vector with $\hat{S}_z$ gives

$$\hat{S}_z|r\rangle = \frac{1}{\sqrt{2}}\hat{S}_z|u\rangle + \frac{1}{\sqrt{2}}\hat{S}_z|d\rangle = \frac{1}{2}\left(\frac{1}{\sqrt{2}}|u\rangle - \frac{1}{\sqrt{2}}|d\rangle\right).$$

But this state-vector is definitely not the state that would result from a measurement of S_z. As we have seen in Example 5.2, that measurement result would be either $+\frac{1}{2}$, leaving the system in state $|u\rangle$, or $-\frac{1}{2}$, leaving it in state $|d\rangle$. Neither of these results is equal to the state represented by the superposition $\frac{1}{\sqrt{2}}|u\rangle - \frac{1}{\sqrt{2}}|d\rangle$. The point is here that measuring S_z destroys any information we may have had about S_x.

Generally, we can say that if a state $|\Psi\rangle$ is not an eigenstate of $\hat{Q}$, then this state of the physical quantity Q is undefined, or meaningless in the sense that quantum theory can assign to it no meaning. In the Example 5.3 $|r\rangle$ is not an eigenstate of $\hat{S}_z$ and thus has no meaning for the observable S_z.

A crucial feature of quantum mechanics compared to classical mechanics is an inherent uncertainty. Uncertainty is not always the case that the result of an experiment is uncertain. If a system is in an eigenstate of an observable, then there is no uncertainty about the result of measuring that observable. But whatever the state, there is always uncertainty about some observable. Uncertainty refers to the spread in the observed values of a physical quantity in non eigenstates. There is no uncertainty in the theory, either in its mathematical formulation, or in its prediction of experimental results.

In our spin example this uncertainty comes about since the spin component, say S_x, changes every time we measure S_z or S_y. Thus no two spin components can be simultaneously measured.

Mathematically, this means that the spin operators do not commute – it makes difference whether we first measure S_x or first measure S_z: $\hat{S}_x\hat{S}_z|\Psi\rangle \neq \hat{S}_z\hat{S}_x|\Psi\rangle$. The shorthand notation for this expression is

$$[\hat{S}_x, \hat{S}_z] := \hat{S}_x\hat{S}_z - \hat{S}_z\hat{S}_x \neq \hat{0},$$

where $[\hat{S}_x, \hat{S}_z]$ is the commutator ([201]) of $\hat{S}_x$ and $\hat{S}_z$ and $\hat{0}$ the null operator. We know that the spin operators are represented by Pauli matrices. We shall see in Section 7.1 that Pauli matrices do not commute: $[\sigma_j, \sigma_k] = 2\mathrm{i}\epsilon_{jkl}\sigma_l$, where ϵ_{jkl} is the Levi-Civita symbol ([202]).

More generally, if a state $|\Psi\rangle$ happens to be an eigenvector of one Hermitian operator – call it $\hat{Q}$ – then it will not be an eigenvector of other operator $\hat{R}$ that do not commute with $\hat{Q}$. Thus, as a rule, if $\hat{Q}$ and $\hat{R}$ do not commute, then there must be uncertainty in one or the other, if not both.

Let us now see how the important quantum operators, the momentum operator $\hat{p}$, the energy operator $\hat{E}$ and the position operator $\hat{x}$, look like ([203]). We have already come upon these operators in Section 5.2 when discussing the Schrödinger equation.

Let us start with momentum. We know that momentum is connected to symmetry under spatial transformations via the Noether's theorem ([204]). On the other hand, the differential operator $\boldsymbol{\nabla}$ generates spatial translations. Therefore, we make the identification $\hat{p} = -\mathrm{i}\hbar\boldsymbol{\nabla}$ ([205]) or simply $\hat{p} = -\mathrm{i}\boldsymbol{\nabla}$, since we assume that $\hbar = 1$ ([206]). Of course,

$$\hat{p} = (\hat{p}_x, \hat{p}_y, \hat{p}_z) = -\mathrm{i}(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}).$$

Analogously, energy is connected to symmetry under temporal translations which are generated by $\mathrm{i}\hbar\frac{\partial}{\partial t}$, or shortly $\mathrm{i}\frac{\partial}{\partial t}$. Consequently, $\hat{E} = \mathrm{i}\frac{\partial}{\partial t}$.

Take note that the (reduced) Planck constant $\hbar$ is introduced in order to ensure correct units of the operators $\hat{p}$ and $\hat{E}$.

Since there is no symmetry connected to the conservation of position, the position operator $\hat{x}$ is just $\boldsymbol{x}$ ([207]).

Now let us compute the commutator $[\hat{p}_a, \hat{x}_b]$, where $a, b = x, y, z$

$$\begin{aligned}
[\hat{p}_a, \hat{x}_b]|\Psi\rangle &= (\hat{p}_a\hat{x}_b - \hat{x}_b\hat{p}_a)|\Psi\rangle \\
&= (-\mathrm{i}\frac{\partial}{\partial a}b + b\mathrm{i}\frac{\partial}{\partial a})|\Psi\rangle \\
&= -(\mathrm{i}\frac{\partial}{\partial a}b)|\Psi\rangle - b\mathrm{i}\frac{\partial}{\partial a}|\Psi\rangle + b\mathrm{i}\frac{\partial}{\partial a}|\Psi\rangle \\
&= -\mathrm{i}\delta_{ab}|\Psi\rangle,
\end{aligned}$$

where δ_{ab} is the Kronecker delta ([208]). In conclusion ([209])

---

[201] Physically, the commutator $[\hat{A},\hat{B}]$ tells us whether measuring observable A affects the measurement of observable B and vice versa.

[202] The Levi-Civita symbol is defined as follows: $\epsilon_{ijk} := \begin{cases} 1 & \text{if } (i,j,k) \in \{(1,2,3),(2,3,1),(3,1,2)\}, \\ 0 & \text{if } i=j \text{ or } j=k \text{ or } k=i, \\ -1 & \text{if } (i,j,k) \in \{(1,3,2),(3,2,1),(2,1,3)\}. \end{cases}$

[203] Notice that these operators act on the Hilbert space $L^2(\mathbb{C})$ of complex *square-integrable* functions defined on spacetime.

[204] Recall that Noether's theorem says that to every continuous symmetry of a theory corresponds a conservation law and vice versa.

[205] For a single particle with no electric charge and spin-0.

[206] The imaginary unit i is introduced in order to make eigenvalues of the operator real. The minus sign is motivated by the Minkowski metric.

[207] All this operator does if it acts on a function $f(x)$ is to multiply it by $\boldsymbol{x}$, i.e., $\hat{x}f(x) = \boldsymbol{x}f(x)$.

[208] The Kronecker delta δ_{ab} is, by definition, zero for $a \neq b$ and one for $a = b$.

[209] This is the simplest case of one particle. In general, the index runs over all degrees of freedom in the system.

(5.41) $$[\hat{p}_a, \hat{x}_b] = -\mathrm{i}\hbar\delta_{ab}.$$

This equation is known as the *canonical commutation relation* and plays a very important role in quantum mechanics. Actually, many textbooks use it as the fundamental postulate of quantum mechanics ([S2]).

In quantum mechanics, average values are called *expectation values*. Suppose we have a probability function for the outcome of an experiment that measures an observable Q. The outcome must be one of $\hat{Q}$'s eigenvalues, λ_i, and the probability function is $P(\lambda_j)$. In bra-ket notation the average value of Q is written $\langle Q\rangle$. From a mathematical point of view, an average is defined by the equation ([210])

$$\langle Q\rangle := \sum_j \lambda_j P(\lambda_j).$$

Suppose that the normalised state of a quantum system is $|\Psi\rangle$ and expand it in the orthonormal basis of eigenvectors of $\hat{Q}$:

$$|\Psi\rangle = \sum_j a_j\,|\lambda_j\rangle.$$

Now let us compute the quantity $\langle\Psi|\hat{Q}|\Psi\rangle$. Recall the meaning of this: First act on $|\Psi\rangle$ with the linear operator $\hat{Q}$. Then, take the inner product of the result with the bra $\langle\Psi|$. Let us do the first step

$$\hat{Q}|\Psi\rangle = \sum_j a_j\hat{Q}\,|\lambda_j\rangle = \sum_j a_j\lambda_j\,|\lambda_j\rangle,$$

because $|\lambda_j\rangle$ are eigenvectors of $\hat{Q}$: $\hat{Q}|\lambda_j\rangle = \lambda_j|\lambda_j\rangle$. The last step is to take the inner product with $\langle\Psi|$. We do that by expanding the bra $\langle\Psi|$ in eigenvectors on the right-hand side, and then using the orthonormality of the eigenvectors. The result is

$$\langle\Psi|\hat{Q}|\Psi\rangle = \sum_j(a_j^* a_j)\lambda_j = \sum_j P(\lambda_j)\,\lambda_j = \langle Q\rangle.$$

We used here the probability principle to identify $a_j^* a_j$ with the probability $P(\lambda_j)$. Consequently, we have a quick rule to compute the averages of observables. Just sandwich the operator corresponding to the observable between the bra and ket representations of the state-vector:

$$\langle Q\rangle = \langle\Psi|\hat{Q}|\Psi\rangle.$$

We now know what a quantum state is. A second crucial ingredient that we need in the quantum framework is something that allows us to determine how a given state evolves in time. That is what we are going to discuss now.

Let us consider a closed system that at time t is in the quantum state $|\Psi\rangle$. To indicate that the state was $|\Psi\rangle$ at the specific time t, we write $|\Psi(t)\rangle$.

The basic dynamical assumption of quantum mechanics is that if we know the state at one time, say at time zero, then the quantum equations of motion tell us what it will be later at time t. The state at time t is given by some operator that we denote $U(t)$, acting on the state at time zero

(5.42) $$|\Psi(t)\rangle = U(t)|\Psi(0)\rangle.$$

The operator U is called *the time-development operator* for the system.

Quantum mechanics imposes a couple of restrictions on U. First, it requires that U is a linear operator. The relationships between states in quantum mechanics are always linear. It goes along with the idea that the state-space is a vector space. It also requires that the operator U preserves distinct states. This implies that if two states are distinguishable (i.e. orthogonal) at time zero they will continue to be distinguishable for all time. We can express this as follows: if $\langle\Psi(0), \Phi(0)\rangle = 0$ then

$$\langle\Psi(t), \Phi(t)\rangle = \langle U(t)\Psi(0), U(t)\Phi(0)\rangle = 0 \text{ for all values of } t.$$

---

[210] The expectation value is the average value of repeated measurements on the **same** state, which is the crucial point.

One can show that this condition is satisfied if $UU^\dagger = I$, where I is the *unit operator* ([211]), called also *identity operator*. An operator that possesses this property is called *unitary*. In physics lingo, the evolution of state vectors with time is unitary.

Notice that the time evolution of the state vector is obviously deterministic. It is exactly like any other linear system state transition. But how does that fit together with the statistical character of our measurement results? As we have seen above in this section, knowing the quantum state $|r\rangle$ does not mean that one can predict the result of a measurement of S_z or S_y with certainty. For this reason, equation (5.42) is not the same as classical determinism. Classical determinism allows us to predict the results of experiments. The quantum evolution of states allows us to compute the probabilities of the outcomes of later experiments. In classical mechanics, there is no real difference between states and measurements. In quantum mechanics, the difference is profound.

If the state vector were the main focus of observational physics, we would say that quantum mechanics is deterministic. But experimental physics is not about measuring the state vector. It is about measuring observables. Nevertheless, between observations, the state of a system evolves in a deterministic way. But something different happens when an observation is made. Measuring an observable Q can have an unpredictable outcome, but after the measurement is made, the system is left in an eigenstate of Q. Which eigenstate? The one corresponding to the outcome of the measurement. But this outcome is unpredictable. So it follows that during an experiment the state of a system jumps unpredictably to an eigenstate of the observable that was measured. This phenomenon is called the *collapse of the wave function*.

In quantum mechanics, we assume that unitary operators are continuous. This means that the state-vector changes smoothly. There is one property that makes continuous operators especially nice to deal with: they can be arbitrarily close to the identity operator I. This means that for small time intervals Δt, a unitary operator U can be written as

$$U(\Delta t) = I + \Delta t H$$

for some operator H. In physics we write by convention

$$U(\Delta t) = I + \mathrm{i}\Delta t H. \tag{5.43}$$

Now, remembering that Hermitian conjugation requires the complex conjugation of coefficients, we find that

$$U^\dagger(\Delta t) = I - \mathrm{i}\Delta t H^\dagger.$$

Since U is unitary, i.e. $UU^\dagger = I$, we infer

$$(I + \mathrm{i}\Delta t H)(I - \mathrm{i}\Delta t H^\dagger) = I.$$

Expanding to first order in Δt, we find $H^\dagger - H = 0$, or in more illuminating form

$$H = H^\dagger.$$

This last equation follows from the unitarity condition and says that H is a Hermitian operator. This has great significance. We can now say that the operator $\widehat{H} := H$ represents an observable and has a complete set of orthonormal eigenvectors and eigenvalues. We will see in a moment that $\widehat{H}$ will become an already familiar object, namely the quantum Hamiltonian operator (see Section 5.2). Its eigenvalues are the values that would result from measuring the energy of a quantum system.

Applying (5.43) we can write equation (5.42) in the form

$$|\Psi(\Delta t)\rangle = (I + \mathrm{i}\Delta t\widehat{H})|\Psi(0)\rangle.$$

Hence

$$\frac{|\Psi(\Delta t)\rangle - |\Psi(0)\rangle}{\Delta t} = \mathrm{i}\widehat{H}|\Psi(0)\rangle.$$

---

[211] The unit operator I is defined by $I|\Psi\rangle = |\Psi\rangle$ for every state $|\Psi\rangle$. Notice that $I = I^\dagger$, i.e. I is Hermitian.

If we take the limit as $\Delta t \to 0$, it becomes the time derivative of the state-vector ([212]):

$$\frac{\partial}{\partial t}|\Psi\rangle = \mathrm{i}\hat{H}|\Psi\rangle. \tag{5.44}$$

This equation has the form of Schrödinger equation (5.3'). However, in order to identify the operator $\hat{H}$ in (5.44) with the Hamiltonian, we have still to correct this equation, because it does not make dimensional sense. After all, in physics, the Hamiltonian is the mathematical object that represents the energy of a system. To resolve this dilemma let us rewrite the equation with Planck's constant inserted in a way that makes it dimensionally consistent:

$$\mathrm{i}\hbar\frac{\partial}{\partial t}|\Psi\rangle = \hat{H}|\Psi\rangle. \tag{5.45}$$

This is exactly Schrödinger equation (5.3'). If we know the Hamiltonian $\hat{H}$, equation (5.45) tells us how the state of an undisturbed system evolves with time.

Because the Hamiltonian operator $\hat{H}$ represents energy, the observable values of energy are just the eigenvalues of $\hat{H}$. Let us call these eigenvalues E_j and the corresponding eigenvectors $|E_j\rangle$. By definition, the relation between $\hat{H}$, E_j, and $|E_j\rangle$ is the eigenvalue equation

$$\hat{H}|E_j\rangle = E_j|E_j\rangle. \tag{5.46}$$

This is the *time-independent Schrodinger equation*.

The quantum theory can be summarized as follows ([H13]):

- The state of a system corresponds to a non-zero vector $|\Psi\rangle$ ([213]) in an abstract 'Hilbert space' ([214]). The vectors $|\Psi\rangle$ and $|\lambda\Psi\rangle$ describe the same state, where $\lambda \neq 0$ is a complex number.
- An observable, such as momentum, is associated with an linear operator that acts on state vectors. When the observable is measured in one of its eigenstates, the corresponding eigenvalue is obtained. However, when the observable is measured in a non-eigenstate, a statistical distribution of eigenvalues is the result. In order to ensure that the eigenvalues are real, it is necessary for the operators to be 'Hermitian'.
- The process of quantization of a classical theory can be achieved through the conversion of the Hamiltonian $H(p, x)$ into an operator $\hat{H}$, facilitated by Heisenberg's commutation relation $[p, x] = -\mathrm{i}\hbar$. This procedure is designated as 'canonical quantization'
- The Hamiltonian is the generator of time evolution, as expressed by the Schrödinger equation (5.45).

The following table provides a concise overview of the principles of quantum formalism.

---

[212] There is nothing special about t = 0, one gets the same result choosing arbitrary time.

[213] It should be noted that the notation $|0\rangle$ does not signify the zero vector, which is represented by 0. The ket $|0\rangle$ is used to denote the vacuum state, that is to say, the lowest energy state of a quantum system, in which no particles are present.

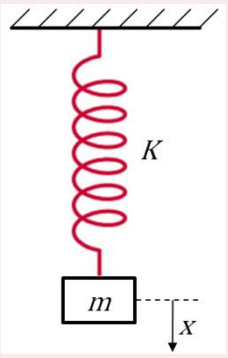
[214] The reader may ask why Hilbert space is used and not, for example, a Banach space. Here are some key reasons:
(1) Hilbert spaces have an inner product, which allows for the definition of orthogonality and projection. This is crucial for quantum mechanics, where the inner product represents the probability amplitude and is used to calculate probabilities. Projections are associated with measurement operators, which correspond to observable quantities. When a measurement is conducted, the state of the system is said to be projected onto an eigenstate of the observable in question.
(2) The spectral theorem applies to operators on Hilbert spaces, allowing for the decomposition of operators into eigenvalues and eigenvectors. This is a prerequisite for understanding the measurement process in quantum mechanics, where observables are represented by operators and their eigenvalues correspond to possible measurement outcomes.
(3) The evolution of quantum states is described by unitary operators in Hilbert space, which preserve the inner product and hence the total probability. This ensures the consistency and conservation of probability in quantum mechanics.
In summary, Hilbert spaces offer the necessary mathematical structure and properties that are essential for the formulation and interpretation of quantum mechanics, making them the natural choice over more general normed spaces.

Physical quantities	Mathematical terms
quantum state	vector in Hilbert space $\mathcal{H}$
observable Q	Hermitian operator $\hat{Q}$:$\mathcal{H} \to \mathcal{H}$
possible results of measurements of Q	eigenvalues of $\hat{Q}$
states associated with unique values of measurements of Q	eigenstates of $\hat{Q}$
time development of a quantum state	unitary operator U

In this section we have discussed ordinary quantum mechanics (QM), rather than quantum field theory (QFT). QFT and QM are different theories, but the difference is not so much in the equations. QFTs are just a special class of quantum mechanical theories where observables are conveniently written as quantum fields or as their functions and functionals. But they describe the same physics – and indeed, the non-relativistic quantum mechanical theories are the limits of QFTs in the case of speeds much lower than the speed of light. When one takes this limit, the Hilbert space of the non-relativistic QM theory is embedded or identified with the Hilbert space of the QFT ([M8]).

QFT is a generalization of the quantum theory to fields. How does it work? It goes something like this ([215]): we take a (classical) field (e.g. electromagnetic field) and decompose it into a (potentially infinite) sum of basic sine wave patterns (harmonic oscillators) using a Fourier transform. Once this decomposition is done, we apply the rules of ordinary quantum mechanics to quantise these harmonic oscillators. This leads to the notion of *second quantisation*, which is used to describe the canonical quantisation of relativistic fields. Each harmonic oscillator will have a lowest energy (ground) state and excitations, stepwise (quantised) higher energy states. It is these stepwise excitations that we recognize as particles. Because interactions create or annihilate excitations, QFT can account for the creation and annihilation of particles, something that ordinary quantum mechanics cannot do. And this very possibility of creation and annihilation of particle-antiparticle pairs (pair creation) is the physical reason for QFT ([T5]).

Example 5.4. ([L1]) The key features of a harmonic oscillator, such as periodic motion and sinusoidal oscillations, can be well illustrated by a mass-spring system. In this system, a mass m is attached to a spring with a spring constant K. When the mass is displaced from its equilibrium position, the spring exerts a restoring force proportional to the displacement x, thereby inducing simple harmonic motion.



The total energy is E is the sum of the kinetic energy and the potential energy

[215] The idea of field quantisation sounds simple but its implementation is difficult and requires a great deal of 'hard-core' mathematics. In QFT fields (typically as the wave functions of matter) are thought of as field operators, in a manner similar to how we considered above operators corresponding to observables (spin, position, momentum, etc.). The key ideas of this method were introduced in 1927 by Paul Dirac, and were developed, most notably, by Fock and Jordan later.

Vladimir Aleksandrovich Fock (1898 – 1974) was a Soviet physicist. Ernst Pascual Jordan (1902 – 1980) was a German theoretical and mathematical physicist ([W11]).

$$E = \frac{p^2}{2m} + \frac{1}{2}Kx^2,$$

where $p = m\dot{x}$ is the momentum. Replacing p by the operator $-i\hbar\nabla$ we get the Schrödinger equation for a harmonic oscillator

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + \frac{1}{2}Kx^2\right)\psi = \mathrm{E}\psi.$$

The solutions to this equation are given by eigenfunctions ψ_n with eigenvalues

$$E_n = (n + \tfrac{1}{2})\hbar\omega, \tag{5.47}$$

where $\omega = \sqrt{K/m}$. The energy levels E_n are arranged in a hierarchical structure, known as a *ladder*. It is noteworthy that when $\mathrm{n} = 0$, the energy is not zero but rather equal to $\hbar\omega/2$. This phenomenon is referred to as the *zero-point energy*.

The Hamiltonian for the harmonic oscillator is expressed as follows:

$$\hat{H} = \frac{1}{2}m\omega^2\hat{x}^2 + \frac{\hat{p}^2}{2m},$$

where the spring constant is expressed as $K = m\omega^2$. The operator $\hat{H}$ can now be written in the following form:

$$\hat{H} = \frac{1}{2}m\omega^2\left[\hat{x}^2 - (\frac{i\hat{p}}{m\omega})^2\right]$$

and we could attempt to apply the formula $a^2 - b^2 = (a - b)(a + b)$ in order to obtain the following result:

$$\hat{H} = \frac{1}{2}m\omega^2\left(\hat{x} - \frac{i\hat{p}}{m\omega}\right)\left(\hat{x} + \frac{i\hat{p}}{m\omega}\right).$$

However, the formula $a^2 - b^2 = (a - b)(a + b)$ is not applicable in this case, since the operators $\hat{x}$ and $\hat{p}$ do not commute. Consequently, the following formula must be employed: $a^2 - b^2 = (a - b)(a + b) - [a, b]$. where $[a, b] = ab - ba$. Applying (5.41) we get

$$\left[\hat{x}, \frac{i\hat{p}}{m\omega}\right] = \frac{i}{m\omega}[\hat{x}, \hat{p}] = -\frac{\hbar}{m\omega}.$$

This implies that

$$\hat{H} = \frac{1}{2}m\omega^2\left[\left(\hat{x} - \frac{i\hat{p}}{m\omega}\right)\left(\hat{x} + \frac{i\hat{p}}{m\omega}\right) + \frac{\hbar}{m\omega}\right]$$

and finally

$$\hat{H} - \frac{\hbar\omega}{2} = \frac{1}{2}m\omega^2\left(\hat{x} - \frac{i\hat{p}}{m\omega}\right)\left(\hat{x} + \frac{i\hat{p}}{m\omega}\right).$$

It thus follows that $\hat{H}$ requires the correction $-\hbar\omega/2$ due to the zero-point energy being subtracted.

We shall now consider the operators $\hat{a}$ and $\hat{a}^\dagger$, defined as follows:

$$\hat{a} = \sqrt{\frac{m\omega}{2\hbar}}\left(\hat{x} + \frac{i}{m\omega}\hat{p}\right), \quad \hat{a}^\dagger = \sqrt{\frac{m\omega}{2\hbar}}\left(\hat{x} - \frac{i}{m\omega}\hat{p}\right).$$

The substitution of the quantities $\hat{a}$ and $\hat{a}^\dagger$ into the Hamiltonian yields the following equation:

$$\hat{H} = \hbar\omega\left(\hat{a}\hat{a}^\dagger + \frac{1}{2}\right).$$

We define the *number operator* $\hat{n}$ by $\hat{n} = \hat{a}^\dagger\hat{a}$. The operator $\hat{n}$ has eigenstates $|n\rangle$ with eigenvalues $n = 0, 1, 2, \ldots : \hat{n}|n\rangle = n|n\rangle$ ([L1]). It can thus be deduced that the eigenstates of $\hat{H}$ will also be of the form $|n\rangle$, with associated eigenvalues of $\hbar\omega(n + 1/2)$. This means that the eigenvalues (5.47) of a harmonic oscillator have been recovered.

The quantity n is used to denote the energy level on the ladder that the system has reached. Alternatively, it may be considered the number of quanta (each of energy $\hbar\omega$) that must have been added to the system when it was in its ground state.

Now let us examine the state defined by $\hat{a}^\dagger|n\rangle$ and apply the number operator to it. Since $[\hat{a}, \hat{a}^\dagger] = \hat{a}\hat{a}^\dagger - \hat{a}^\dagger\hat{a} = 1$, we obtain

$$\hat{n}\hat{a}^\dagger|n\rangle = \hat{a}^\dagger\hat{a}\hat{a}^\dagger|n\rangle = \hat{a}^\dagger\big(1 + \hat{a}^\dagger\hat{a}\big)|n\rangle = \hat{a}^\dagger|n\rangle + \hat{a}^\dagger\hat{n}|n\rangle = \hat{a}^\dagger|n\rangle + \hat{a}^\dagger n|n\rangle = (n+1)\hat{a}^\dagger|n\rangle.$$

This demonstrates that the state $\hat{a}^\dagger|n\rangle$ is an eigenstate of $\hat{H}$, with an eigenvalue that is one higher than the state $|n\rangle$. For this reason $\hat{a}^\dagger$ is called the *creation* (or *raising*) operator, which increases the number of particles in the state by one.

In a similar manner, it can be demonstrated that the state defined by $\hat{a}|n\rangle$ is an eigenstate of $\hat{H}$, but with an eigenvalue that is one lower than the state $|n\rangle$. For this reason $\hat{a}$ is called the *annihilation* (or *lowering*) operator, which decreases the number of particles in the state by one.

In summary, the operator $\hat{a}^\dagger$ can be regarded as the entity responsible for the creation of a quantum of energy $\hbar\omega$, and the subsequent movement of the oscillator up one rung of the energy ladder. Conversely, its adjoint, $\hat{a}$, acts to annihilate a quantum of energy $\hbar\omega$ and move the oscillator down one rung of the ladder. It is important to note that these quanta of energy behave like particles: the application of these operators results in the addition and subtraction of particles.

We end this section with the following comment. Comparing equations (5.3') and (5.45) we treated a wave function as though it were the state vector. This ambiguous use of terminology can be confusing. It becomes less confusing when we realize that a wave function can represent a state vector. To see it let us expand a state vector $|\Psi\rangle$ in a specific basis of eigenvectors $|\varphi_j\rangle$ corresponding to a given quantum operator

$$|\Psi\rangle = \sum_j a_j\,|\varphi_j\rangle. \tag{5.48}$$

Since the state-vector $|\Psi\rangle$ changes with time and the basis vectors $|\varphi_j\rangle$ do not, it follows that the coefficients a_j must also depend on time: $a_j = a_j(t)$. We can also consider a_j as a function of both t and $\boldsymbol{x}$: $a_j = a_j(t, \boldsymbol{x})$. In general, different operators have different eigenstates. This means that we can expand a general state vector in terms of different eigenstates corresponding to a different operator. Notice that operators which commute have a common set of eigenstates.

For example, if we are interested in the energy of the system, we expand our general state vector in terms of energy eigenvectors $|E_j\rangle$. By (5.33) each coefficient in the expansion is equal to the inner product of $|\Psi\rangle$ with one of the basis vectors: $a_j = \langle\varphi_j|\Psi\rangle$. Moreover, as we have seen in Example 5.2, the coefficients a_j tell us the probability to measure a given result. The set of possible outcomes need not be discrete. In the case of a continuous set of possible outcomes, the general state in (5.48) is expressed as an integral rather than a sum. Also, the set of discrete coefficients $a_j = a_j(t, \boldsymbol{x})$ is then replaced by a function $\psi = \psi(\mathrm{t}, \boldsymbol{x})$. This function is called the *wave function*.

The form of the wave function depends on which observables we choose to focus on. That is because calculations for two different observables rely on different sets of basis vectors. According to the basic probability principle of quantum mechanics, the squared magnitude of the wave function is the probability for the observables to have specific values (= eigenvalues). An example of a wave function is the function $\psi(\mathrm{x})$ that we get by expanding a state vector in terms of position eigenstates

$$|\Psi\rangle = \int dx\; \psi(x)\,|x\rangle.$$

Let us return now to our main topic.

6. Classical Yang-Mils theory [216]

The history of the Yang-Mills theory is long and complicated (see e.g. [O1], [O2], [Q2]). In 1935, Yukawa [217] suggested that the strong nuclear forces might be mediated by massive mesons. Yukawa proposed in his paper [Y4] a theory of the strong interaction between a proton and a neutron, and also considered its possible extension to neutron β-decay. He built his theory by analogy with electromagnetism, postulating a new field of force with an associated new field quantum, analogous to the photon. In doing so, he showed how particles interact by exchanging virtual quanta, which mediate the force. ([A1])

Although Yukawa's ideas did not work as a theory of strong interactions [218], his approach was profound, and the ideas have broad and lasting validity.

In 1954, Chen Ning 'Frank' Yang [219] and Robert Mills [220] proposed a mathematical scheme that might be useful for the strong interaction, which (among other things) binds protons and neutrons together in the nucleus [221]. Yang and Mills wondered if they could find some symmetry among particles that would dictate their interactions, and found a promising-looking candidate called 'isotopic spin', first described by Werner Heisenberg [222] in 1932. Just as the phase of the wave function in electromagnetism can be shifted arbitrarily in spacetime because the interaction with the electromagnetic field A_μ cancels out the effect of the alteration (see Section 5.4), so Yang and Mills proposed to do the same for isotopic spin, hypothesising the existence of a '$\boldsymbol{B}_\mu$ field' to counteract the change ([C7]).

The SU(2) gauge theory they found did not work for this purpose, since (as we shall see in Section 7.5) what was needed was an SU(3) theory of quarks and gluons which came only 20 years later [223]. The SU(2) gauge theory of isotopic spin they were considering ultimately did find a role in the electroweak part of the Standard model, but this idea got started only after the symmetry properties of the weak interactions became clear later in the 1950s. Schwinger and his student Glashow were among the first to work on this idea, with the correct theory not appearing until 1967 after the role of the Higgs mechanism was understood ([W13]).

In 1953, Pauli could have been the author of the seminal discovery of gauge theories which constitute the basis of our current understanding of nature at short distances. He was interested in a six-dimensional theory of general relativity along the lines suggested by Th. Kaluza [224] and O. Klein in five dimensions. He compactified two extra dimensions into two-dimensional sphere, which led him to SU(2) gauge theory. However, non-Abelian gauge bosons remain massless, and at that time the only massless fields known to physicists were photons and the neutrino

---

216 See [A1] and [Y1] as a general reference for this chapter.

217 Hideki Yukawa (1907 – 1981) was a Japanese theoretical physicist and the first Japanese Nobel laureate (1949) for his prediction of the pi meson.

218 Yukawa assumed that the nucleons and his quantum (later identified with the pion) were point-like, but in fact both nucleons and pions are quark composites.

219 Yang Chen-Ning or Chen-Ning Yang (1922 –), also known as C. N. Yang or by the English name Frank Yang, is a Chinese theoretical physicist. Yang is the only citizen of the People's Republic of China who has won the Nobel Prize in Physics (1957).

220 Robert Laurence Mills (1927 – 1999) was an American physicist.

221 See also R. Shaw [S7]. Ronald Shaw, a post-graduate student of Abdus Salam at Cambridge, working under the influence of Schwinger invented gauge field theory in his doctoral thesis, independently of (and almost simultaneously with) Yang & Mills. The maths was very elegant, but it appeared to have no application in nature (because of the mass problem). Therefore Shaw and his supervisor decided not to submit the work for publication. When Salam heard of the Y-M paper, he advised Shaw to publish his results (which he did not).

222 Werner Heisenberg (1901 – 1976) was a German theoretical physicist and one of the key pioneers of quantum mechanics. 1932 Nobel Prize in Physics.

223 SU(n) denotes the special unitary group of degree n. It is the group of n×n unitary matrices with determinant 1 (see Section 7.1 for details).

224 Theodor Franz Eduard Kaluza (1885 – 1954) was a German mathematician and physicist.

postulated by Pauli in 1930 ([225]) ([S8]), and not yet discovered in 1953 ([226]). Moreover, such a particle would mediate a long range force instead of the short-range force of the strong and weak interactions.

In late 1953, Pauli's enthusiasm began to wane. *"If one tries to formulate field equations one will always obtain vector mesons with rest mass zero. One could try to get other meson fields - pseudoscalars with positive rest mass. But I feel that is too artificial"* ([P1]). Because of his super-high requirements for his own work in physics, Pauli put on hold publication on his theory ([227]). Pauli applied extremely high criteria of 'cleanliness' both to his own works and to those of other theoretical physicists and was not afraid of open conflicts in those cases when he saw gaps or imperfections in the line of reasoning. ([228])

We begin with a discussion of the idea of isospin.

6.1. Isospin and SU(2) symmetry

Because like charges repel, it is remarkable that the atomic nucleus stays together. After all, the protons are all positively charged and are repelled from each other electrically. To hold these particles so closely together, physicists hypothesised a new force, the *strong force*, strong enough to overcome the electric repulsion of the protons. It must be strongest only at short distances (about 10^{-15} m – see Chapter 2), and then it must fall off rapidly, for protons are repelled electrically unless their separation is that small. Neutrons must also experience it because they are bound to the nucleus as well ([H16]).

Physicists spent several decades trying to understand the strong force; it was one of the principal problems in physics in the mid-twentieth century. When the neutron was discovered in 1932, it was natural to assume that this was a composite particle consisting of a proton and an electron. Heisenberg used the neutron-as-proton-plus-electron idea to develop an early theory of proton–neutron interactions in the nucleus. He hypothesized that the proton and neutron bind together in the nucleus by exchanging an electron between them, the proton turning into a neutron and the neutron turning into a proton in the process. ([B1])

Heisenberg proposed in [H3] that the proton and neutron could be two states of a single *nucleon* ([H16]). These states are differentiated by an internal property that can have two values, $+{}^{1}/_{2}$ and $-{}^{1}/_{2}$, in analogy with the (true) spin of a particle such as the electron. This new

---

[225] The neutrino was postulated by Pauli to explain how beta decay could conserve energy, momentum, and angular momentum (spin). Pauli, who was unwilling to give up the conservation laws, conjectured the existence of a new particle. This was a neutral particle of spin ½ with a mass "*not larger than 0.01 proton mass*", as Pauli suggested in a famous letter sent on December 4, 1930, to nuclear physicists who were holding a meeting in Tübingen, Germany (see Section 7.4). He proposed that each electron in the nucleus was accompanied by one of the new particles, which he provisionally named neutrons. Pauli let the matter rest, presenting his idea publicly at the Solvay Conference in October 1933 held in Brussels. The word 'neutrino' entered the scientific vocabulary through Enrico Fermi, who used it during a conference in Paris in July 1932 and at the Solvay Conference in 1933, where Pauli also employed it. The name (the Italian equivalent of 'little neutral one') was jokingly coined by Edoardo Amaldi during a conversation with Fermi at the Institute of Physics of via Panisperna in Rome, in order to distinguish this light neutral particle from heavy neutron ([W11]).

[226] The discovery of neutrino took two decades to accomplish, since the neutrino can pass through light-years of matter without interacting. It was first observed in 1956 by a group led by Clyde L. Cowan and Frederick Reines of Los Alamos National Laboratory. In 1995, Frederick Reines was awarded the Nobel Prize in Physics for the discovery of the neutrino. (Clyde Cowan died in 1974.)

[227] In December 1953 Pauli wrote in a letter to Pais [P1]: "*So this leads to some rather unphysical `shadow particles'*." It was clear to him that the gauge bosons had to be massless. This must have been the reason why he did not publish anything. ([S13])

[228] Pauli had already long been recognized as one of the major figures in twentieth-century physics, not only because of his own contributions, but also because of his critical judgments – which could be quite sharp, but nearly always to the point – of others' work. He was known as the conscience of twentieth-century physics. Pauli's critical mind could not bear unsound results, incomplete works, or hand-waving arguments. It was important that he applied the same high criteria to his own results. Very instructive in this respect is the story of his last work with Heisenberg which remained unpublished ([P1]).

property is called *isotopic spin*, or *isospin* for short ([229]), and the nuclear binding force is said to exhibit *isospin symmetry*. The nucleon has then two allowed states (the proton p and the neutron n) which are not distinguished by the nuclear force. Converting a neutron into a proton is then equivalent to 'rotating' the spin of the neutron in an 'isospin-space' (which has just two dimensions, up and down) from spin-down to spin-up. ([B1])

The families of similar particles are known as isospin multiplets: two-particle families are called doublets, three-particle families are called triplets, and so on. The doublet (n, p) is grouped together in an isospin multiplet with total isospin $I = {}^1/_2$, with projection $I_3 = +{}^1/_2$ for the proton and $I_3 = -{}^1/_2$ for the neutron ([230]).

The rest energies of the proton and neutron are almost the same: $m_\mathrm{p} = 938.28\ \mathrm{MeV/c^2}$, $m_\mathrm{n} = 939.57\ \mathrm{MeV/c^2}$ ([231]). Following the mass-energy equivalence of special relativity $E = mc^2$, this mass equivalence can be viewed as an energy degeneracy ([232]) of the underlying interactions. Quite generally in quantum mechanics, we know that whenever we have a set of states which are degenerate in energy (or mass) there is no unique way of specifying the states: any linear combination of some initially chosen set of states will do just as well, provided the normalisation conditions on the states are still satisfied. ([A1])

This single near coincidence of the masses m_p and m_n was enough to suggest to Heisenberg that, as far as the strong nuclear forces were concerned (electromagnetism being negligible by comparison), the two states could be regarded as truly degenerate, so that any arbitrary linear combination of neutron ψ_n and proton ψ_p wave functions would be entirely equivalent, as far as this force was concerned, for a single 'neutron' or single 'proton' wave function. This hypothesis became known as the *charge independence of nuclear forces* ([233]). Thus redefinitions of neutron and proton wave functions could be allowed, of the form

$$\psi_\mathrm{p} \to \psi'_\mathrm{p} = \mathrm{u}\psi_\mathrm{p} + \mathrm{v}\psi_\mathrm{n}, \qquad \psi_\mathrm{n} \to \psi'_\mathrm{n} = \mathrm{w}\psi_\mathrm{p} + \mathrm{z}\psi_\mathrm{n}, \tag{6.1}$$

where u, v, w and z are complex numbers. ([A1])

If the proton and the neutron are to be viewed as two linearly independent states of the same particle, it is natural to represent them in terms of a two component vector, analogous to the spin-up $\begin{bmatrix}1\\0\end{bmatrix}$ and spin-down $\begin{bmatrix}0\\1\end{bmatrix}$ states of a spin-½ system. The analogy can be brought out by introducing the *two-component nucleon isospinor* (see [A1; Chapter 12] for more details)

$$\psi^{(1/2)} := \begin{bmatrix}\psi_\mathrm{p}\\ \psi_\mathrm{n}\end{bmatrix} = \psi_\mathrm{p}\begin{bmatrix}1\\0\end{bmatrix} + \psi_\mathrm{n}\begin{bmatrix}0\\1\end{bmatrix}. \tag{6.2}$$

In the proton-neutron doublet $\psi^{(1/2)}$, ψ_p is the amplitude for the nucleon to have 'isospin up' and ψ_n is that for it to have 'isospin down'. Linear combinations of ψ_n and ψ_p correspond to quantum-mechanical superpositions of the two states, and the particle then looks sometimes like proton and sometimes like neutron.

---

[229] The isotopic spin was first introduced by E. Wigner [W9] and B. Cassen and E. U. Condon [C2]. One should however distinguish this isospin from the *weak isospin*, which (as we shall see in Section 7.4) is an attribute of leptons as well as of quarks in the context of weak interactions and, hence, physically quite distinct from the isospin considered here.

[230] This projection is analogous to the spin operator $\hat{S}_z$ in Example 5.3.

[231] We now believe that that this small difference is due the near equality of the up and down quarks.

[232] *Degenerate* is used in quantum mechanics to mean 'of equal energy'. The number of different states of equal energy is called the *degree of degeneracy* or just *degeneracy*.

[233] More accurately, Heisenberg supposed that strong interaction physics remains invariant if one exchanges the proton and the neutron. Note that this symmetry is considerably weaker than isospin symmetry, in which one transforms the proton and the neutron into linear combinations of each other. In 1936, B. Cassen and E. U. Condon, and, independently, G. Breit and E. Feenberg, proposed that Heisenberg's exchange symmetry be generalized to isospin symmetry. ([Z2])

Despite this formal analogy between the isospin and spin-up and spin-down states of a spin-½ system, it is important to be clear, however, that the two cases are quite distinct. In particular, even though both the proton and the neutron have (true) spin-½, the transformations (6.1) leave the (true) spin part of their wave functions completely untouched ([234]). Thus, the isospin has nothing to do with the true spin of the particles, i.e. it is not a spin in physical space with a corresponding angular momentum. This is a 'spin' in an abstract space with no associated angular momentum. ([A1])

Heisenberg's proposal, then, was that the physics of strong interactions between nucleons remained the same under the transformation (6.3): in other words, a symmetry was involved. It, therefore, seems that all protons and neutrons could be interchanged and the strong interaction would hardly be altered. If the electromagnetic forces could somehow be turned off, the isospin symmetry would be exact; in reality it is only approximate. ([A1])

The weaknesses of the Heisenberg's theory were exposed in experiments performed just a few years later. Because protons do not possess a 'stuck-on' electron, the electron-exchange model did not allow for any kind of interaction between protons. In contrast, experiments showed that the strength of the interaction between protons is comparable to that between protons and neutrons. Despite the shortcomings of the theory, Heisenberg's electron-exchange model held at least a grain of truth. The exchange of electrons was abandoned, but the concept of isospin was retained. ([B1])

If isospin invariance would be an exact symmetry then it is a matter of convention which component of $\psi^{(1/2)}$ would correspond to the proton and which one to the neutron. lf one insists on being able to define this convention at any spacetime point separately, then one is led to the construction of a gauge field theory based on local isospin transformations – this is the heuristic argument that motivated Yang and Mills to attempt the construction of the gauge theory of SU(2) ([W12]). They asked the question whether the reference frame used to define the isospins could vary from point to point. If so, the information that a particle produced at a given spacetime point was e.g. a proton would be meaningless for a different observer, unless there existed a way to compare their two frames. This is a role of the gauge field, very much as in electrodynamics relative phases of charged fields at different points make sense only when compared via the electromagnetic potential A_μ. ([I2])

Equations (6.1) can be written in terms of $\psi^{(1/2)}$ as

$$\psi^{(1/2)} \to \psi^{(1/2)\prime} = \mathrm{U}\psi^{(1/2)}, \tag{6.3}$$

where U is the complex 2×2 matrix

$$\mathrm{U} := \begin{bmatrix} \mathrm{u} & \mathrm{v} \\ \mathrm{w} & z \end{bmatrix}. \tag{6.4}$$

The matrix (6.4) depends on four arbitrary complex numbers or, alternatively, on eight real parameters. However, it is subject to certain restrictions ([235]) and these reduce the number of free parameters in U to three ([A1]). Consequently, instead of arbitrary matrices one considers in (6.3) only so-called *special unitary* 2×2 matrices.

'Special' simply means that U has unit determinant $\det(\mathrm{U}) = 1$. A matrix U is called *unitary* provided $\mathrm{U}^\dagger\mathrm{U} = \mathrm{U}\mathrm{U}^\dagger = \mathrm{I}$, where I is the identity matrix and $\mathrm{U}^\dagger := (\mathrm{U}^T)^*$. For example, for the matrix (6.4) we have

$$\mathrm{U}^\dagger = \begin{bmatrix} \mathrm{u}^* & \mathrm{w}^* \\ \mathrm{v}^* & \mathrm{z}^* \end{bmatrix}.$$

Every special unitary 2×2 matrix U has the form

$$\mathrm{U} = \begin{bmatrix} \mathrm{u} & \mathrm{v} \\ -\mathrm{v}^* & \mathrm{u}^* \end{bmatrix}$$

---

[234] These transformations just mix the two components of the isospinor (i.e. the proton and neutron) inside the doublet $\psi^{(1/2)}$.

[235] For example, the transformation (6.3) must preserve the normalisation of $\psi^{(1/2)}$. This implies that U has to be unitary.

where u, v are complex numbers with the property $|u|^2 + |v|^2 = 1$.

Notice that for complex vectors we want the inner product of a vector with itself to be a real number because by definition this should result in the squared length of the vector. A complex number would make little sense as the length of the vector. Therefore, the inner product uu, where $u = (u_1, u_2, \ldots u_n) \in \mathbb{C}^n$, is defined with additional complex conjugation $uu := u^\dagger u = uu^\dagger = \sum_{i=1}^n u_i^* u_i = \sum_{i=1}^n |u_i|^2 \in \mathbb{R}$ (cf. Section 5.9).

In order for the physical predictions to be unchanged by a transformation U, it must preserve the length of state vectors and this is the case if and only if U is unitary, i.e. $U^\dagger U = I$.

The set of all special unitary 2×2 matrices with ordinary matrix multiplication form a group called SU(2) ([236]). Of course, it is a non-Abelian group because the matrix multiplication is in general not commutative (cf. Example 3.3). Thus the isospin symmetry is governed by a non-Abelian group SU(2) 'rotating' components of the doublet $\psi^{(1/2)}$ into each other in abstract isospin space $\mathbb{C}^2$ ([237]). This enables one to utilize what is already known about the SU(2) symmetry group from the study of angular momentum.

Now let us show that the transformation (6.3) is of essentially the same mathematical form as the (global) phase transformation $\psi \to \psi' = e^{i\alpha}\psi$.

A complex matrix H is called *Hermitian* if $H = H^\dagger$, i.e. if it is equal to its own conjugate transpose. The *trace* of a square matrix H, denoted Tr(H), is defined to be the sum of elements on the main diagonal (from the upper left to the lower right) of H. H is called *traceless* if $\mathrm{Tr}(H) = 0$. Thus if H is a 2×2 traceless Hermitian matrix then it must have the form ([238])

$$H = \begin{bmatrix} a & b - ic \\ b + ic & -a \end{bmatrix} \tag{6.5}$$

where a, b, c are real numbers. Putting $a = \alpha_3/2$, $b = \alpha_1/2$ and $c = \alpha_2/2$ we can write (6.5) as follows

$$H = \frac{1}{2}\boldsymbol{\alpha}\cdot\boldsymbol{\tau} \tag{6.6}$$

where $\boldsymbol{\alpha}$ stands for the three real numbers $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ and $\boldsymbol{\tau}$ for three matrices $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3)$ which are just the familiar Pauli spin matrices (5.23)

$$\tau_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \tau_2 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \quad \tau_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \tag{6.7}$$

here called 'tau' in order to distinguish them from the mathematically identical 'sigma' matrices which are associated with the real spin degree of freedom. ([A1])

Let us show that (6.6) holds. We have

$$\frac{1}{2}\boldsymbol{\alpha}\cdot\boldsymbol{\tau} = \frac{1}{2}(\alpha_1\tau_1 + \alpha_2\tau_2 + \alpha_3\tau_3) = \frac{1}{2}\left(\begin{bmatrix} 0 & \alpha_1 \\ \alpha_1 & 0 \end{bmatrix} + \begin{bmatrix} 0 & -i\alpha_2 \\ i\alpha_2 & 0 \end{bmatrix} + \begin{bmatrix} \alpha_3 & 0 \\ 0 & -\alpha_3 \end{bmatrix}\right) =$$

$$= \frac{1}{2}\begin{bmatrix} \alpha_3 & \alpha_1 - i\alpha_2 \\ \alpha_1 + i\alpha_2 & -\alpha_3 \end{bmatrix} = \begin{bmatrix} a & b - ic \\ b + ic & -a \end{bmatrix} = H.$$

It is known that every special unitary matrix U has the form e^{iH} ([239]) for some traceless Hermitian matrix H (see Section 7.1). Consequently, if U belongs to SU(2) then it has the form

---

[236] The S stands for 'special', U stands for 'unitary' and 2 for 2×2 matrices.

[237] Prior to isospin, the symmetries of physics (translation invariance, rotation invariance, Lorentz invariance, and so on) were confined to the spacetime. Heisenberg's proposal led to the discovery of a vast internal space, the ongoing exploration of which has been a central theme of fundamental physics for close to a hundred years now ([Z1]).

[238] In general, n×n Hermitian matrices contains $n^2 - 1$ adjustable (real) constants.

[239] In mathematics, the matrix exponential is a matrix function on square matrices analogous to the ordinary exponential function e^x. If V is a square matrix then the exponential of V, denoted by e^V or exp(V), is the matrix given by the power series $e^V := \sum_{k=0}^{\infty} \frac{1}{k!} V^k$ where $V^0 := I$ is defined to be the identity matrix with the same dimensions as V. The power series always converges, e.g. with respect to the norm $||V|| = \sqrt{\sum_{i,j=1}^n |V_{ij}|^2}$, where $V = [V_{ij}]$.

(6.8) $$U = e^{i\frac{1}{2}\boldsymbol{\alpha}\cdot\boldsymbol{\tau}}$$

and we can write the transformation (6.3) as

(6.9) $$\psi^{(1/2)} \to \psi^{(1/2)\prime} = e^{i\frac{1}{2}\boldsymbol{\alpha}\cdot\boldsymbol{\tau}}\psi^{(1/2)}.$$

Now it is clear that the isospin transformation (6.9) is a generalization of the global phase transformation $\psi \to \psi' = e^{i\alpha}\psi$, except that

- there are now three 'phase angles' α_i, one for each Pauli matrix τ_i, and
- there are non-commuting matrix operators (the τ's) appearing in the exponent. ([A1])

The $\boldsymbol{\tau}$ matrices play an important role since they determine the forms of the three (linearly) independent SU(2) transformations. They are called the *generators* of SU(2) transformations (more precisely, they are generators of the algebra $\mathfrak{su}(2)$ – see Section 7.1 for details).

Since the elements of SU(2) are parameterized by two complex numbers (6.5), with the sum of their squared length equal to one, they are vectors of length one in $\mathbb{R}^4$, when we identify $\mathbb{C}^2 = \mathbb{R}^4$. Just as U(1) could be identified as a space with the unit circle in $\mathbb{C} = \mathbb{R}^2$ (see Section 5.5), SU(2) can be identified with the unit three-sphere S^3 in $\mathbb{R}^4$.

Finally, let us notice that the internal SU(2) symmetry (6.3)

$$\psi^{(1/2)}(t, \boldsymbol{x}) \to \psi^{(1/2)\prime}(t, \boldsymbol{x}) = U\psi^{(1/2)}(t, \boldsymbol{x})$$

is a global one because the matrix U does not depend on $(t, \boldsymbol{x})$.

The stage is now set for the discussion of the Yang-Mills paper [Y1].

6.2. The Yang-Mills paper

On October 1st 1954 the 32 years old Chen Ning Yang and somewhat younger Robert Mills published the paper [Y1] *"Conservation of Isotopic Spin and Isotopic Gauge Invariance"* in which they asked the following question: Could one replace global isospin rotations (6.3) by local (spacetime dependent) ones? This would mean that the matrix U would depend on the points of spacetime, just like the gauge generator $\Lambda(t, \boldsymbol{x})$ ([240]) in electromagnetism:

$$(t, \boldsymbol{x}) \to U(t, \boldsymbol{x}) = \begin{bmatrix} u(t, \boldsymbol{x}) & v(t, \boldsymbol{x}) \\ w(t, \boldsymbol{x}) & z(t, \boldsymbol{x}) \end{bmatrix}.$$

Yang and Mills [Y1]: "*We define 'isotopic gauge' as an arbitrary way of choosing the orientation of the isotopic spin axes at all spacetime points, in analogy with the electromagnetic gauge which represents an arbitrary way of choosing the complex phase factor of a charged field at all space-time points. We then propose that all physical processes (not involving the electromagnetic field) be invariant under an isotopic gauge transformation,* $\psi' = U\psi$*, where* U *represents a space-time dependent isotopic spin rotation.*"

The local SU(2) transformation for an isospin doublet wave function

(6.10) $$\psi^{(1/2)}(t, \boldsymbol{x}) \to U(t, \boldsymbol{x})\psi^{(1/2)}(t, \boldsymbol{x}) = e^{i\frac{1}{2}\boldsymbol{\alpha}(t,\boldsymbol{x})\cdot\boldsymbol{\tau}}\psi^{(1/2)}(t, \boldsymbol{x})$$

leads however to a problem similar to that in electromagnetism. To write down field equations for protons and neutrons, one needs the derivatives of these fields. The way these derivatives transform under a gauge transformation (6.10) implies that there will be terms containing the gradients $\nabla_\mu U$ of the matrices. To make the theory gauge-invariant, these gradients would have to be cancelled out, and in order to do that, Yang and Mills replaced the derivative ∇_μ by a covariant derivative D_μ as was done in electromagnetism (see Section 5.4). ([H10])

---

[240] See (5.11).

In the electromagnetic case, the covariant derivative is $D_\mu = \nabla_\mu + \mathrm{i}qA_\mu$ which under a local U(1) phase transformation (5.14) transforms as a matter field ψ with charge q (see Section 5.2)

$$D_\mu\psi \to \mathrm{V}(\mathrm{t}, \boldsymbol{x})D_\mu\psi(\mathrm{t}, \boldsymbol{x}), \tag{6.11}$$

where $\mathrm{V}(\mathrm{t}, \boldsymbol{x}) = \mathrm{e}^{\mathrm{i}q\Lambda(\mathrm{t},\boldsymbol{x})}$ is a local U(1) transformation corresponding to the generator q – see Example 7.7 below ([241]). The gauge transformation (4.19) of A_μ can be then written as

$$A_\mu \to \mathrm{V}A_\mu\mathrm{V}^\dagger - \frac{\mathrm{i}}{q}(\nabla_\mu\mathrm{V})\mathrm{V}^\dagger. \tag{6.12}$$

Indeed, we have $\mathrm{V}A_\mu\mathrm{V}^\dagger = \mathrm{e}^{\mathrm{i}q\Lambda(\mathrm{t},\boldsymbol{x})}\mathrm{e}^{-\mathrm{i}q\Lambda(\mathrm{t},\boldsymbol{x})}A_\mu = A_\mu$ and $-\frac{\mathrm{i}}{q}(\nabla_\mu\mathrm{V})\,\mathrm{V}^\dagger =$
$-\frac{\mathrm{i}}{q}(\nabla_\mu\mathrm{e}^{\mathrm{i}q\Lambda(\mathrm{t},\boldsymbol{x})})\mathrm{e}^{-\mathrm{i}q\Lambda(\mathrm{t},\boldsymbol{x})} = -\frac{\mathrm{i}}{q}\mathrm{e}^{-\mathrm{i}q\Lambda(\mathrm{t},\boldsymbol{x})}\mathrm{e}^{\mathrm{i}q\Lambda(\mathrm{t},\boldsymbol{x})}\nabla_\mu\mathrm{i}q\Lambda(\mathrm{t}, \boldsymbol{x}) = \nabla_\mu\Lambda(\mathrm{t}, \boldsymbol{x})$.
Hence (6.12) amounts to (4.19): $A_\mu \to A_\mu - \nabla_\mu\Lambda$.

Property (6.11) ensures the gauge covariance of wave equations in the U(1) case. By analogy, the key property one demands for a SU(2) covariant derivative D_μ is that the quantity $D_\mu\psi^{(1/2)}$ should transform like $\psi^{(1/2)}$ in (6.10). It means that

$$D_\mu\psi^{(1/2)} \to \mathrm{U}(\mathrm{t}, \boldsymbol{x})D_\mu\psi^{(1/2)} = e^{\mathrm{i}\frac{1}{2}\boldsymbol{\alpha}(\mathrm{t},\boldsymbol{x})\cdot\boldsymbol{\tau}}D_\mu\psi^{(1/2)} \tag{6.13}$$

must hold. Yang and Mills defined D_μ as follows

$$D_\mu := \mathrm{I}_2\nabla_\mu - \mathrm{i}g\frac{1}{2}\boldsymbol{\tau}\cdot\boldsymbol{b}_\mu(\mathrm{t}, \boldsymbol{x}), \tag{6.14}$$

where I_2 is the 2×2 identity matrix and $\boldsymbol{\tau}$ stands for three Pauli matrices $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3)$. The $\boldsymbol{b}_\mu(\mathrm{t}, \boldsymbol{x})$ denotes three independent vector fields $\boldsymbol{b}_\mu = (b_\mu^1, b_\mu^2, b_\mu^3)$ ([242]) generalizing the single electromagnetic gauge field A_μ ([243]). The parameter g is coupling strength, analogous to the electromagnetic charge q ([244]).

As we have already mentioned in Section 4.4, when a global invariance is generalized to a local one, the existence of a new 'compensating' fields is entailed, interacting in a specified way.

The term $\boldsymbol{\tau}\cdot\boldsymbol{b}_\mu$ in (6.14) is then the 2×2 matrix

$$\boldsymbol{\tau}\cdot\boldsymbol{b}_\mu = \tau_1 b_\mu^1 + \tau_2 b_\mu^2 + \tau_3 b_\mu^3 = \begin{bmatrix} b_\mu^3 & b_\mu^1 - \mathrm{i}b_\mu^2 \\ b_\mu^1 + \mathrm{i}b_\mu^2 & -b_\mu^3 \end{bmatrix},$$

the $(\mathrm{t}, \boldsymbol{x})$-dependence of the b_μ's is understood ([A1]). Notice that the derivative D_μ is also a 2×2 matrix

$$D_\mu = \begin{bmatrix} \nabla_\mu & 0 \\ 0 & \nabla_\mu \end{bmatrix} - \mathrm{i}g\frac{1}{2}\begin{bmatrix} b_\mu^3 & b_\mu^1 - ib_\mu^2 \\ b_\mu^1 + ib_\mu^2 & -b_\mu^3 \end{bmatrix},$$

so it can act on a two-component isospinor $\psi^{(1/2)}$ ([245]). Putting $B_\mu := \frac{1}{2}\boldsymbol{\tau}\cdot\boldsymbol{b}_\mu$ we can write (6.14) in the form

---

[241] The parameter q labels the representation to which the matter field ψ belongs - see Section 7.1.

[242] Please note that despite the μ subscript, $\boldsymbol{b}_\mu$ is not a 4-vector. In this shorthand the index i of b_μ^i is hidden. One must presume its presence from the context. Each vector b_μ^i, $i = 1, 2, 3$, has four real components $(b_0^i, b_1^i, b_2^i, b_3^i)$.

[243] Now there are three fields because the $\mathfrak{su}(2)$ algebra has three generators – see Section 7.1 and Example 7.8.

[244] The coupling constant ('gauge coupling parameter') g determines the strength of the interaction with the fields $\boldsymbol{b}_\mu$. This is analogous to the fine-structure constant $\alpha = e^2/4\pi \approx 1/137$, which quantifies the strength of the electromagnetic interaction between fundamental charged particles. Strictly speaking, neither g nor α is a constant.

[245] The notation $D_\mu\psi^{(1/2)}$ should be understood as $\begin{bmatrix}(D_\mu\psi)_1 & (D_\mu\psi)_2\end{bmatrix}^T$, where $(D_\mu\psi)_k = \nabla_\mu\psi_k - \mathrm{i}g\frac{1}{2}\sum_{l=1}^{2}(\boldsymbol{\tau}\cdot\boldsymbol{b}_\mu)_{kl}\psi_l$, $k = 1, 2$, $\psi_1 = \psi_\mathrm{p}$ and $\psi_2 = \psi_\mathrm{n}$.

$$D_\mu = \mathrm{I}_2\nabla_\mu - \mathrm{i}gB_\mu$$

which is analogous to the covariant derivative $D_\mu = \nabla_\mu + \mathrm{i}qA_\mu$ in the electromagnetic case. This time, however, $(\mathrm{t}, \boldsymbol{x}) \to B_\mu(\mathrm{t}, \boldsymbol{x})$ is not a 4-vector field but consists of four matrix-valued fields.

In order to ensure that equation (6.13) is satisfied, Yang and Mills defined the transformation law for SU(2) gauge fields $\boldsymbol{b}_\mu$ as follows

$$\frac{1}{2}\boldsymbol{\tau}\cdot\boldsymbol{b}'_\mu = \mathrm{U}\frac{1}{2}\boldsymbol{\tau}\cdot\boldsymbol{b}_\mu\mathrm{U}^\dagger - \frac{\mathrm{i}}{g}(\nabla_\mu\mathrm{U})\mathrm{U}^\dagger, \tag{6.15}$$

where $\mathrm{U} = \mathrm{U}(\mathrm{t}, \boldsymbol{x}) \in \mathrm{SU}(2)$. Writing (6.15) in the form ([246])

$$B'_\mu = \mathrm{U}B_\mu\mathrm{U}^\dagger - \frac{\mathrm{i}}{g}(\nabla_\mu\mathrm{U})\mathrm{U}^\dagger$$

we can see that this (more complicated) transformation is similar to the gauge transformation (6.12) of the electromagnetic gauge field A_μ, the U(1) transformation V being replaced by the SU(2) transformation U ([247]) ([A1]).

In analogy to the procedure of obtaining electromagnetic field-strength tensor (4.7), the SU(2) field strength tensor, is given by

$$\boldsymbol{F}_{\mu\nu} := \nabla_\mu\boldsymbol{b}_\nu - \nabla_\nu\boldsymbol{b}_\mu + g\boldsymbol{b}_\mu\times\boldsymbol{b}_\nu. \tag{6.16}$$

One might try to define the field strength tensor of as $\nabla_\mu\boldsymbol{b}_\nu - \nabla_\nu\boldsymbol{b}_\mu$, but this quantity does not have the required transformation property under U ([248]). On the other hand, the field strength tensor (6.16) transforms as needed

$$\boldsymbol{F}_{\mu\nu} \to \mathrm{U}\boldsymbol{F}_{\mu\nu}\mathrm{U}^\dagger.$$

The quantity $\boldsymbol{F}_{\mu\nu}$ has three components $F^1_{\mu\nu}, F^2_{\mu\nu}, F^3_{\mu\nu}$ where

$$F^i_{\mu\nu} = \nabla_\mu b^i_\nu - \nabla_\nu b^i_\mu + g(\boldsymbol{b}_\mu\times\boldsymbol{b}_\nu)^i. \text{ ([249])}$$

Notice that each $F^i_{\mu\nu}$, $i = 1, 2, 3$, is a 4×4 matrix $[F^i_{\mu\nu}]$, $\mu, \nu = \mathrm{t}, \mathrm{x}, \mathrm{y}, \mathrm{z}$.

One can show that analogous to (5.12)

$$[D_\mu, D_\nu] := D_\mu D_\nu - D_\nu D_\mu = \frac{1}{2}\mathrm{i}g\boldsymbol{\tau}\cdot\boldsymbol{F}_{\mu\nu}. \tag{6.17}$$

Since the tensor $\boldsymbol{F}_{\mu\nu}$ arises from the commutator of two gauge-covariant derivatives, it is itself SU(2) gauge covariant, i.e. it transforms under local SU(2) transformations in the expected way ([250]). However, $\boldsymbol{F}_{\mu\nu}$ contains a non-linear term $g\boldsymbol{b}_\mu\times\boldsymbol{b}_\nu$, which makes the Yang-Mills equations a lot more complicated than Maxwell's equations. In particular, the presence of the

---

[246] In order to ensure that a global transformation U conserves the physical properties of the system, the field B_μ should transform as $B_\mu \to \mathrm{U}B_\mu\mathrm{U}^\dagger$. Local transformations require additionally the term $\mathrm{i}/g(\nabla_\mu\mathrm{U})\mathrm{U}^\dagger$ – see Section 7.2.

[247] We take $g > 0$, while the electron charge is negative. This is the origin of some apparent sign differences in the definitions for the U(1) and the non-Abelian case SU(2) ([M1]).

[248] Yang had searched for this tensor without success since his student days in 1947. As he recalls ([H13]):
"*I was clearly focusing on a very important problem. Unfortunately the mathematical calculations always ended in more and more complicated formulas and total frustration. It was only in 1953–1954, when Bob Mills and I revisited the problem and tried adding quadratic terms to the field strength $F_{\mu\nu}$ that an elegant theory emerged. For Mills and me it was many years later that we realized the quadratic terms were in fact natural from the mathematical point of view.*" ([Y3])

[249] For example, $F^1_{xy} = -\frac{\partial}{\partial x}b^1_y + \frac{\partial}{\partial y}b^1_x - g(b^2_x b^3_y - b^3_x b^2_y)$, because $(\boldsymbol{b}_\mathrm{x}\times\boldsymbol{b}_\mathrm{y})^1$ being the first component of the product $(b^1_x, b^2_x, b^3_x)\times(b^1_y, b^2_y, b^3_y)$ is equal to $b^2_x b^3_y - b^3_x b^2_y$. Usually the following notation is used: $(\boldsymbol{b}_\mathrm{x}\times\boldsymbol{b}_\mathrm{y})^i = \epsilon_{ijk}b^j_\mu b^k_\nu := \sum_{j,k=1}^3 \epsilon_{ijk}b^j_\mu b^k_\nu$ where ϵ_{ijk} denotes the Levi-Civita symbol. Recall that b^i_μ is a vector with components $(b^i_t, b^i_x, b^i_y, b^i_z)$.

[250] By SU(2) gauge covariant we mean that it transforms according to the adjoint representation Ad of the group SU(2) – see Section 7.2 for details.

gauge coupling constant g in this term means that the fields b_μ^1, b_μ^2, b_μ^3 themselves carry SU(2) 'charge' and act as sources for the field strength. Consequently, these gauge fields will necessarily interact with themselves and therefore a gauge theory of non-Abelian fields alone has non-trivial interactions and is not a free theory.

This is profoundly different from the electromagnetic case, where the gauge field A_μ for the photon is of course uncharged (photon has no electric charge) and the third term in (6.16) is absent for A_μ in (4.7). ([A1])

In order to derive the field equations for the $\boldsymbol{b}_\mu$ fields, Yang and Mills introduced the following Lagrangian density

$$\mathcal{L}_{YM} = -\frac{1}{4}\boldsymbol{F}_{\mu\nu}\boldsymbol{F}^{\mu\nu} + \bar{\psi}\gamma_\mu(\nabla^\mu - \mathrm{ig}\boldsymbol{\tau}\cdot\boldsymbol{b}^\mu)\psi - m\,\bar{\psi}\psi. \tag{6.18}$$

where $\psi = \psi^{(1/2)}$ and both the 'isospinor' components of $\psi^{(1/2)}$ are four-component Dirac spinors ([251]). Writing (6.18) in the form

$$\mathcal{L}_{YM} = \bar{\psi}(\gamma_\mu\nabla^\mu - m)\psi - \frac{1}{4}\boldsymbol{F}_{\mu\nu}\boldsymbol{F}^{\mu\nu} - \mathrm{g}\bar{\psi}\gamma_\mu\boldsymbol{\tau}\cdot\boldsymbol{b}^\mu\psi, \tag{6.19}$$

we can see similarities with the Dirac-Lagrangian density (5.30). We have here the 'free' term $\bar{\psi}(\gamma_\mu\nabla^\mu - m)\psi$, the 'kinetic term' $\mathcal{L}_{YM}^{kinetic} = -\frac{1}{4}\boldsymbol{F}_{\mu\nu}\boldsymbol{F}^{\mu\nu}$ and the 'interaction term' $\mathrm{g}\bar{\psi}\gamma_\mu\boldsymbol{\tau}\cdot\boldsymbol{b}^\mu\psi$. This interaction term couples the gauge fields $\boldsymbol{b}_\mu$ with the isospin (fermionic) field ψ.

In the derived field equations (Euler-Lagrange equations) appears the quantity $\boldsymbol{j}_\mu = \mathrm{ig}\bar{\psi}\gamma_\mu\boldsymbol{\tau}\psi$ which the authors use to define the *isotopic spin current density*

$$\boldsymbol{j}_\mu^{YM} = \boldsymbol{j}_\mu + 2\mathrm{g}\boldsymbol{b}_\nu\times\boldsymbol{F}_{\mu\nu}.$$

This density satisfies the equation of continuity

$$\nabla_\mu\boldsymbol{j}_\mu^{YM} = 0.$$

It implies that $\boldsymbol{j}_\mu^{YM}$ is conserved. This corresponds to the law of conservation of electric charge in electrodynamics. Note that $\boldsymbol{j}_\mu^{YM}$ contains the gauge fields itself. That is, each gauge field carries charge, and acts as its own source. In contrast, the electromagnetic field is neutral, and does not have intrinsic self-interaction ([H13])

In the further part of the paper the fields $\boldsymbol{b}_\mu$ are quantised, which results in complex equations. In the last section, the authors formulate the following statement regarding the quanta (= bosons) of the fields $\boldsymbol{b}_\mu$:

„*The quanta of the* $\boldsymbol{b}$ *field clearly have spin unity and isotopic spin unity. We know their electric charge too because all the interactions that we propose must satisfy the law of the conservation of electric charge, which is exact. The two states of the nucleon, namely proton and neutron, differ by charge unity. Since they can transform into each other through the emission of a* $\boldsymbol{b}$ *quantum, the latter must have three charge states with charges of* $\pm\mathrm{e}$ *and* 0".

Thus the Yang-Mills theory predicts and describes a new type of three spin-1 particles (i.e. three bosons associated with the Yang-Mills $\boldsymbol{b}_\mu$ fields) that transmit a force not unlike the electromagnetic force. Just as electromagnetic gauge invariance requires the existence of massless photons, so Yang-Mills symmetry invokes three massless vector gauge bosons, two with opposite electric charges, one neutral.

---

[251] Reminder: γ_μ are the Dirac matrices (5.24) and $\bar{\psi} := \psi^\dagger\gamma_0$. Moreover, the scalar term $\boldsymbol{F}_{\mu\nu}\boldsymbol{F}^{\mu\nu}$ is actually equal to the sum $\sum_{i=1}^{3} F_{\mu\nu}^i F^{i\,\mu\nu}$. Notice that 4×4 matrices γ_μ act on the four components of the Dirac spinors ψ_p and ψ_n, whereas the SU(2) transformation (being a 2×2 matrix) mixes the two Dirac spinors inside the doublet $\psi = \begin{bmatrix}\psi_\mathrm{p}\\ \psi_\mathrm{n}\end{bmatrix}$.
Recall that $\boldsymbol{b}^\mu := \eta^{\nu\mu}\boldsymbol{b}_\mu = (\eta^{\nu\mu}b_\mu^1, \eta^{\nu\mu}b_\mu^2, \eta^{\nu\mu}b_\mu^3)$, where $\eta^{\nu\mu}$ is the Minkowski metric

The main problem of the Yang-Mills theory is that it is not possible to add a mass term for the gauge fields of the form $\frac{1}{2}m^2\boldsymbol{b}_\mu \cdot \boldsymbol{b}^\mu$ to the Lagrangian density $\mathcal{L}_{YM}$, since such a term would not be invariant under the gauge transformations (6.15) of the $\boldsymbol{b}_\mu$ fields. Hence, just as in the U(1) (electromagnetic) case, the $\boldsymbol{b}_\mu$ quanta of this theory have to be massless. But Yang and Mills actually wanted to develop a theory for describing strong interactions of mesons (pions, kaons, vector mesons). However, these forces only act over very short distances, i.e. their exchange particles must be massive (vector) bosons. The authors were well aware of the problem with the mass of the $\boldsymbol{b}_\mu$ bosons. They write:

„*We next come to the question of the mass of the* **b** *quantum, to which we do not have a satisfactory answer.* (...) *In electrodynamics, by the requirement of electric charge conservation, it is argued that the mass of the photon vanishes. Corresponding arguments in the* **b** *field case do not exist even though the conservation of isotopic spin still holds. We have therefore not been able to conclude anything about the mass of the* **b** *quantum. A conclusion about the mass of the* **b** *quantum is of course very important in deciding whether the proposal of the existence of the* **b** *field is consistent with experimental information*."

They made no further progress, and turned their attentions elsewhere.

Because of the problem with the massless bosons, the conviction prevailed that the Yang-Mills theory did not describe a physical reality ([252]). The attempt by Yang and Mills to construct a 'gauge theory of nuclear forces' failed, among other things, for the reason that the nuclear interaction is merely 'phenomenological', i.e. it is far removed from the actual fundamental force, the strong interaction. But that was not known in 1954. That is why this great idea of gauge invariance as the generator of fundamental interactions remained dormant for so long. The Yang-Mills theory did serve to bring gauge theory to the general attention of theorists, but several developments had to come.

7. Comeback of Yang-Mills theory

Today, the phrase 'Yang-Mills' can be used in two ways: to refer to the specific model proposed by Yang and Mills in 1954, or as shorthand for any non-Abelian gauge theory that is relevant to contemporary physics ([C7]). Gauge theories are fundamental to the Standard Model because the model is based on a generalization of the Yang-Mills proposal, the first non-abelian gauge theory dealing with particle symmetries.

However, the massless nature of the Yang-Mills field was a serious stumbling block to applying Yang-Mills theory to the other forces, for the weak and nuclear forces are short range and many of the particles are massive. Hence these phenomena did not appear to be associated with long-range fields describing massless particles.

In the 1960s and 1970s, physicists overcame these obstacles to the physical interpretation of non-Abelian gauge theory. In the case of the weak force, this was accomplished by the Glashow-

---

[252] On February 23, 1954, Yang was invited by Oppenheimer to return to Princeton for a few days and to give a seminar on his joint work with Mills. Here, Yang's report [Y2]:

"*Pauli was spending the year in Princeton, and was deeply interested in symmetries and interactions. Soon after my seminar began, when I had written down on the blackboard,* $(\nabla_\mu - \mathrm{i}gB_\mu)\psi$, *Pauli asked, 'What is the mass of this field* B_μ*?' I said we did not know. Then I resumed my presentation, but soon Pauli asked the same question again. I said something to the effect that that was a very complicated problem, we had worked on it and had come to no definite conclusions. I still remember his repartee: 'That is not sufficient excuse.' I was so taken aback that I decided, after a few moments' hesitation to sit down. There was general embarrassment. Finally Oppenheimer said, 'We should let Frank proceed.' I then resumed, and Pauli did not ask any more questions during the seminar. I don't remember what happened at the end of the seminar. But the next day I found the following message:*
'February 24, Dear Yang, I regret that you made it almost impossible for me to talk with you after the seminar. All good wishes. Sincerely yours, W. Pauli.'"

Salam-Weinberg electroweak theory with the gauge group SU(2)×U(1), which avoided the massless problem by introducing an additional *Higgs field* – see Section 7.4.

The solution to the problem of massless Yang-Mills fields for the strong interactions has a completely different nature. That solution did not come from adding fields to Yang-Mills theory, but by discovering a remarkable property of the quantum Yang-Mills theory itself. This property is called *asymptotic freedom*. Roughly speaking, this means the decrease of the effective interaction strength at high energies or short distances. This made it possible to describe the strong force by a non-Abelian gauge theory in which the gauge group is SU(3) – see Section 7.5.

In order to proceed further, we have to now discuss how group formalism is used to construct gauge theories in more mathematical detail.

7.1. An interlude on group representations

When studying a physical system, the main theoretical tool that physicists use to formulate a mathematical description of the system is the investigation of its symmetries. These symmetries correspond to transformations of the system that leave the physics invariant. Such transformations generally form an abstract group (the symmetry group), and then one can use the language of group theory to describe the physical system. This is why group theory is ubiquitous in modern theoretical physics. But in fact, in physics what we are interested in is how groups act on something: an object, a theory. This is what representation theory is about: it represents the elements of a group as acting on something. ([B5])

In particular, the mathematical framework of quantum mechanics is closely related to what mathematicians describe as the *representation theory of Lie groups* ([253]). These representations are the tools needed to describe all fundamental particles, each representation for a different kind of elementary particle. The representations tell us what types of fundamental particles exist in nature.

A detailed explanation of the representation theory is beyond the scope of this article, but in this section we will introduce some of its main ingredients. Consequently, the discussion of Lie groups and their representations is focused on specific examples, not the general theory.

This section is rather mathematical and formal. The effort will pay, however, since an understanding of this group theoretical approach provides a deeper understanding the construction of gauge theories relevant for fundamental interactions in the Standard Model. By studying group theory and representation theory, we will learn a lot about physical entities. We will see how the spin of a particle arises naturally in terms of representations and how particles in the Standard Model transform according to representations of the gauge groups U(1), SU(2) and SU(3).

In mathematics, the *general linear group of degree* n over the set of complex numbers $\mathbb{C}$, written as GL(n, $\mathbb{C}$), is the set of n×n invertible matrices with complex entries, with the group operation that of ordinary matrix multiplication:

$$\mathrm{GL}(n, \mathbb{C}) = \mathrm{GL}(n) := \{U \in \mathbb{C}^{n\times n} : U \text{ is invertible }\}.$$

An n×n matrix U is called *invertible* if there exists an n×n matrix U^{-1} such that $UU^{-1} = U^{-1}U = I_n$, where I_n denotes the n×n identity matrix ([254]). The set GL(n) forms a group, because the product of two invertible matrices is again invertible, and the inverse of an invertible matrix is invertible, with identity matrix I_n as the identity element of the group. Note that for $n \geq 2$ the group GL(n) is non-Abelian, since matrix multiplication is then non-

[253] See [H0], [S3] and [W14] as a general reference for this section.

[254] A matrix U is invertible if and only if its determinant det(U) is not equal to 0.

commutative (see Example 3.3). The elements of GL(n) can be thought of as linear transformations acting on the n-dimensional vector space $\mathbb{C}^n$ ([255]).

We may similarly define GL(n, $\mathbb{R}$) to be the group of all n×n invertible matrices with real entries. Of course, GL(n, $\mathbb{R}$) is contained in GL(n, $\mathbb{C}$).

More generally, if V is a complex vector space then GL(V) denotes the group of bijective (i.e. one-to-one and onto) linear transformations L:V → V. If V is n-dimensional then choosing a basis for V we can identify GL(V) with GL(n, $\mathbb{C}$) (respectively with GL(n, $\mathbb{R}$), if V is real).

Notice that the set GL(n) with the usual matrix addition $[a_{ij}]+[b_{ij}] = [a_{ij}+b_{ij}]$ and scalar multiplication $\lambda[a_{ij}] = [\lambda a_{ij}]$ is not a vector space because the sum of two invertible matrices need not be invertible.

A *Lie group* is, roughly speaking, a continuous group, that is, a group described by several real parameters which can take continuous values ([256]). The minimal number of such parameters is the *dimension* of the Lie group.

In this article, we consider the Lie groups that are the most common in physics: the matrix Lie groups, i.e. Lie groups realised as groups of matrices. More precisely, a *matrix Lie group* is a closed subgroup G of GL(n) ([257]). The condition that G be a closed subgroup, as opposed to merely a subgroup, should be regarded as a technicality – all for us relevant subgroups of GL(n) have this property.

Of course, GL(n) is itself a matrix Lie group. Since GL(n) is a subset of $\mathbb{C}^{n\times n}$, it has $2n^2$ real parameters and hence its dimension dim[GL(n)] is equal to $2n^2$. Here, by 'dimension' we mean the necessary number of degrees of freedom it takes to parameterize GL(n). We do not mean the dimension of the matrices, which is just n.

Another example is the *unitary group* of degree n, denoted U(n). It is the group of n×n (complex) unitary matrices, with the group operation of matrix multiplication. Recall (see Section 6.1) that a matrix U is unitary provided $U^\dagger U = UU^\dagger = I$, where $U^\dagger := (U^T)^* = (U^*)^T$ is the conjugate transpose and I is the identity matrix. Consequently

$$U(n) := \{U \in \mathbb{C}^{n\times n} : U^\dagger U = UU^\dagger = I\} = \{U \in \mathbb{C}^{n\times n} : U^\dagger = U^{-1}\}.$$

The unitary group U(n) is a subgroup of the general linear group GL(n, $\mathbb{C}$) since every unitary matrix U is invertible with $U^{-1} = U^\dagger$. The unitarity condition acts as a constraint and reduces the number of real parameters of U(n) to n^2, so $\dim[U(n)] = n^2$.

In Section 6.1 we encountered the special unitary group SU(2). More generally, we can consider the group SU(n) of special unitary n×n matrices with the group operation being matrix multiplication:

$$SU(n) := \{U \in U(n) : \det(U) = 1\}.$$

---

[255] If $U = [U_{ij}]$ is an n×n matrix and $u = (u_1, \ldots, u_n) \in \mathbb{C}^n$, then $L(u) := Uu^T = (\sum_{j=1}^n U_{1j}u_j, \sum_{j=1}^n U_{2j}u_j, \ldots, \sum_{j=1}^n U_{nj}u_j)$ describes a linear transformation L: $\mathbb{C}^n \to \mathbb{C}^n$. Indeed, $L(c_1u_1 + c_2u_2) = c_1L(u_1) + c_2L(u_2)$, for $u_1, u_2 \in \mathbb{C}^n$ and $c_1, c_2 \in \mathbb{C}$.

[256] Lie groups play a central role in physics. They are named after Norwegian mathematician Sophus Lie (1842–1899). A standard definition of a Lie group is as a smooth manifold, with group laws given by smooth (infinitely differentiable) maps. It means that a Lie group is a group which is also a manifold. In particular, there is a small neighbourhood around the group identity 1 which looks like a piece of $\mathbb{R}^n$, with n the dimension of the group.
If G and H are Lie groups then a mapping r: G → H is called a *homorphism* if r(ab) = r(a)r(b) for all a, b ∈ G. If in addition, r is one-to-one and onto, then r is called an *isomorphism*. If there exists an isomorphism from G to H, then G and H are said to be *isomorphic*. Two groups which are isomorphic should be thought of as being (for all practical purposes) the same group. It is known that not every Lie group is isomorphic to a matrix Lie group – see e.g. [H0] for details.

[257] G is *closed* provided if (U_k) is a sequence in G such that $U_k \to U$ for some U ∈ GL(n), then U ∈ G. We say, (U_k) converges to U ∈ GL(n) if $||U_k - U|| \to 0$, as $k \to \infty$, where $||V|| = \sqrt{\sum_{i,j=1}^n |V_{ij}|^2}$ and $V = [V_{ij}] \in GL(n)$. It is a general fact that any closed subgroup of a Lie group is a Lie group.

The constraint $\det(U) = 1$ implies that $\dim[SU(n)] = n^2 - 1$. Of course,

$$SU(n) \subseteq U(n) \subseteq GL(n, \mathbb{C}).$$

Moreover, both U(n) and SU(n) are compact ([258]) Lie groups.

Recall (see Section 6.1) that if H is an n×n matrix then we define the *exponential* of H, denoted e^H or exp(H), by the usual power series

$$e^H := \sum_{k=0}^{\infty} \frac{1}{k!} H^k,$$

where $k! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot k$, $H^0 := I$ is the identity matrix, and H^k is the repeated matrix product of H with itself.

The exponential of a matrix plays a crucial role in the theory of Lie groups. The exponential enters into the definition of the Lie algebra of a Lie group (see below) and is the mechanism for passing information from the Lie algebra to the Lie group. Since many computations can be done much more easily at the level of the Lie algebra, the exponential is indispensable ([H0]).

For a matrix Lie group one defines the corresponding *Lie algebra* as the collection of matrices that give an element of the group when exponentiated ([259]). More explicitly, for a Lie group $G \subseteq GL(n)$, the *Lie algebra* of G, denoted $\mathfrak{g}$ ([260]), is the set of all n×n matrices H such that $e^{itH} \in G$ for all $t \in \mathbb{R}$, together with an operation, called the *Lie bracket* $[\cdot, \cdot] : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ that tells us how we can combine these matrices:

$$\mathfrak{g} := \{H \in GL(n) : e^{itH} \in G \text{ for every } t \in \mathbb{R}\}. \text{ ([261])}$$

The Lie bracket

$$[H, K] := HK - KH \tag{7.1}$$

is commonly referred to as the *commutator* of H and K.

It is important to notice that the elements of the Lie algebra are matrices, but the multiplication of two Lie algebra elements does not need to be an element of the Lie algebra. However, the commutator always is: $H, K \in \mathfrak{g} \Rightarrow [H, K] \in \mathfrak{g}$. Consequently, the natural (non-associative) product operation of the Lie algebra is not ordinary matrix multiplication, but the Lie bracket $[\cdot, \cdot]$.

The set $\mathfrak{g}$ with the usual matrix addition and scalar multiplication is a real linear space, and when equipped with the product operation $[\cdot, \cdot]$, it becomes an algebra. It is known that $\mathfrak{g}$ is a real linear subspace of GL(n) ([262]).

The Lie algebra is a very important tool in studying Lie groups. On the one hand, Lie algebras are simpler than Lie groups, because the Lie algebra is a linear space. Thus we can understand

---

[258] $G \subseteq GL(n, \mathbb{C})$ is *compact* provided every sequence (U_k) in G has a convergent subsequence $U_{k_n} \to U$ for some $U \in G$, i.e. $||U_{k_n} - U|| \to 0$, as $n \to \infty$. Clearly, compact subsets are closed.

[259] In mathematics, an algebra A over $\mathbb{C}$ or $\mathbb{R}$ is a vector space equipped with a bilinear product $[\cdot, \cdot] : A \times A \to A$. Thus, an algebra is an algebraic structure consisting of a set together with operations of multiplication, addition and scalar multiplication and satisfying the axioms implied by 'vector space' and 'bilinear' ([W11]).

[260] The Lie algebra which belongs to a group G is conventionally denoted by the corresponding 'Fraktur' letter $\mathfrak{g}$.

[261] We introduce the additional 'i' in the exponent, in order to use Hermitian matrices, which guarantees that we get real numbers as predictions for experiments in quantum mechanics. This is the physics convention to take $U = \exp(iH)$ with H Hermitian, and the mathematics convention is to take $U = \exp(H)$ with H anti-Hermitian, i.e. $H = -H^\dagger$. Notice that if H is Hermitian then iH is anti-Hermitian.

[262] The Lie algebra $\mathfrak{g}$ of a Lie group G is isomorphic to the tangent (vector) space T_1G at the identity element 1 of the group G. An important point is that the Lie algebra is a real vector space, although it is a subspace of a space of complex matrices. For example, $\mathfrak{u}(n)$ is a real vector space, but it is NOT a space of real n×n matrices.
In physics, Lie algebra is frequently referred to as the space of 'infinitesimal group elements', which actually connects the concept of Lie algebra back to its original definition as the tangent space.

much about Lie algebras just by using tools of linear algebra theory. On the other hand, the Lie algebra of a Lie group contains much information about that group. Thus many questions involving a Lie group G can be answered by considering a similar but easier problem for the Lie algebra $\mathfrak{g}$ ([H0]).

Notice that while a group G determines the Lie algebra $\mathfrak{g}$, the Lie algebra does not determine the group. It means that if two Lie groups are isomorphic then their Lie algebras are also isomorphic ([263]), but not otherwise. For example, SU(2) and SO(3) ([264]) are different (i.e. not isomorphic) groups with the same Lie algebra (that is, the Lie algebras $\mathfrak{su}(2)$ and $\mathfrak{so}(3)$ are isomorphic ([W14])).

Let $G \subseteq U(n)$ be a Lie group and $\mathfrak{g}$ its Lie algebra. We show that elements of $\mathfrak{g}$ are Hermitian matrices. Indeed, if $H \in \mathfrak{g}$ and $t \in \mathbb{R}$ then by definition $e^{itH} \in G \subseteq U(n)$, i.e. the matrix e^{itH} is unitary. Recall that a matrix U is unitary if and only if $U^\dagger = U^{-1}$. Thus $(e^{itH})^\dagger = (e^{itH})^{-1} = e^{-itH}$. But by taking Hermitian conjugate term-by-term

$$(e^{itH})^\dagger = \left(\sum_{k=0}^{\infty} \frac{1}{k!}(itH)^k\right)^\dagger = \sum_{k=0}^{\infty} \frac{1}{k!}(-itH^\dagger)^k = e^{-itH^\dagger}$$

we obtain that $e^{-itH^\dagger} = e^{-itH}$ or equivalently $H = H^\dagger$, i.e. H is a Hermitian matrix. And conversely, if H is Hermitian then e^{itH} is unitary. Consequently, the Lie algebra $\mathfrak{g}$ consists of Hermitian matrices. In particular, the Lie algebra $\mathfrak{u}(n)$ of U(n) is the set of all Hermitian n×n matrices.

Now, since every Lie algebra is a linear space, it has a basis. In this context, we call the basis elements *generators*. The number of generators defines the (algebraic) *dimension* of the Lie algebra $\dim(\mathfrak{g})$. We have $\dim(G) = \dim(\mathfrak{g})$.

For a Lie group $G \subseteq U(n)$, the algebra $\mathfrak{g}$ is generated by Hermitian n×n matrices T_j, $j = 1, 2, ..., m$, where $m = \dim(\mathfrak{g})$. As a consequence, every element H of $\mathfrak{g}$ is given by a linear combination of Hermitian generators:

$$H = \alpha_j T_j := \sum_{j=1}^{m} \alpha_j T_j , \tag{7.2}$$

where α_j, $j = 1, 2, ..., m$, are real coefficients.

For a Lie group G, an important question is that of whether the exponential function maps the Lie algebra $\mathfrak{g}$ back onto the entire Lie group G. In general, the answer is 'no'. However, if a Lie group G is either simply connected ([265]) or $G = U(n)$ then each group element $U \in G$ has at least one element H of $\mathfrak{g}$ such that $U = e^{iH}$

$$U = e^{i\sum_{j=1}^{m} \alpha_j T_j}. \tag{7.3}$$

Thus, in this case, we have $G = \{e^{iH} : H \in \mathfrak{g}\}$, i.e. $\mathfrak{g}$ determines G.

The reason why Hermitian generators are important in gauge theory is because gauge fields are associated with generators of the Lie algebra of the gauge symmetry group under

---

[263] If $\mathfrak{g}$ and $\mathfrak{h}$ are Lie algebras then a linear mapping $r:\mathfrak{g} \to \mathfrak{h}$ is called a *Lie algebra isomorphism* if r is bijective (i.e. one-to-one and onto) and $r([x, y]) = [r(x), r(y)]$ for all $x, y \in \mathfrak{g}$. Two algebras which are isomorphic should be thought of as being (for all practical purposes) the same algebra.

[264] SO(3), the 3-dimensional *special orthogonal* group, is a collection of orthogonal 3×3 matrices U with real entries. *Orthogonal* means that the columns of the matrix U have to be orthogonal to one another (i.e. $U^T U = I$), and the word *special* means the matrices have to have determinant 1. The group SO(3) is also known as the three-dimensional *rotation group*. It is the collection of rotations of three-dimensional space that preserve one distinguished point. Rotations are linear transformations of $\mathbb{R}^3$ and can therefore be represented by matrices once a basis of $\mathbb{R}^3$ has been chosen.

[265] A Lie group G is *simply connected* if it is path-connected and every closed path in G may be deformed continuously to a point in G. G is *path-connected* if there is a path joining any two points in G. In particular, SU(n) is simply connected, whereas U(n) is only path-connected.

consideration. For example, as we shall see later (Example 7.9), every generator of the Lie algebra $\mathfrak{su}(3)$ is associated with one of the eight gluon fields.

Be aware that in the physics literature one does not always distinguish clearly between a Lie group and its Lie algebra. For example, the elements of some chosen basis for the Lie algebra $\mathfrak{g}$ are simply called 'generators of the group G'. It can be confusing because 'generators' form a basis for the linear space $\mathfrak{g}$ and hence they need not be elements of G (G need not be a linear space, after all). However, generators of $\mathfrak{g}$ are related to elements of G by the exponential map (7.3).

Example 7.1. The Lie algebra $\mathfrak{u}(n)$ of U(n) is the set of all complex Hermitian n×n matrices

$$\mathfrak{u}(n) = \{H \in GL(n) : H = H^\dagger\}.$$

Since $\dim[\mathfrak{u}(n)] = n^2$, the algebra $\mathfrak{u}(n)$ has n^2 generators.

The simplest example of a Lie group is the group of rotations of the plane, with elements parameterised by a single number, the angle of rotation α. It is useful to identify such group elements with unit vectors $e^{i\alpha}$ in the complex plane.

The group is then denoted U(1), since such complex numbers can be thought of as 1×1 unitary matrices. Since any Hermitian 1×1 matrix is real, we may identify it with a real number. Thus $\mathfrak{u}(1)$ is just the real line (the 1-dimensional vector space $\mathbb{R}$) and has exactly one generator, the 1-dimensional vector $T = 1$, or any other real number ≠ 0.

Example 7.2. Let $G = SU(n) := \{U \in U(n) : \det(U) = 1\}$. Recall that the trace Tr(H) of an n×n matrix $H = [h_{ij}]$ is defined to be the sum of elements on the main diagonal of H

$$\mathrm{Tr}(H) := \sum_{i=1}^{n} h_{ii}.$$

It is known ([H0, Th. 3.10]) that $\det(e^H) = e^{\mathrm{Tr}(H)}$. Thus if $\mathrm{Tr}(H) = 0$, then $\det(e^{itH}) = 1$ for all real t. On the other hand, if $e^{it\mathrm{Tr}(H)} = \det(e^{itH}) = 1$ for all t, then $\mathrm{Tr}(H) = 0$. Thus the Lie algebra $\mathfrak{su}(n)$ of SU(n) is the set of all complex Hermitian n×n matrices with vanishing trace ([266])

$$\mathfrak{su}(n) = \{H \in GL(n) : H = H^\dagger \text{ and } \mathrm{Tr}(H) = 0\}.$$

Since $\dim[\mathfrak{su}(n)] = n^2 - 1$, the algebra $\mathfrak{su}(n)$ has $n^2 - 1$ generators being traceless Hermitian n×n matrices.

Example 7.3. One of the most important examples in physics is the Lie group SU(2). For example, transitions in any 2-state quantum mechanical system are described by this group (c.f. Section 6.1).

The Lie algebra $\mathfrak{su}(2)$ of SU(2) consists of the Hermitian traceless 2×2 matrices and has three generators, because $\dim[\mathfrak{su}(2)] = \dim[SU(2)] = 2^2 - 1 = 3$. A possible set of generators is formed of the Pauli spin matrices – see (5.23)

$$\sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad \sigma_2 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \qquad \sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Conventionally, one rather takes as generators the matrices $T_j := \frac{1}{2}\sigma_i$, $j = 1, 2, 3$. Accordingly, every Hermitian traceless 2×2 matrix can be written as a real linear combination of these matrices. Thus by (7.3), every matrix $U \in SU(2)$ is of the form

$$U = e^{i\sum_{j=1}^{3} \alpha_j T_j} = e^{i\boldsymbol{\alpha}\cdot\boldsymbol{T}},$$

where $\boldsymbol{\alpha}$ stands for three real numbers $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ and $\boldsymbol{T} = (T_1, T_2, T_3)$.

---

[266] This is how the det(U) = 1 condition of SU(n) turns into a constraint on the generators of the algebra $\mathfrak{su}(n)$.

Notice that we cannot allow complex linear combinations of generators. Take, for example, the matrix iT_1. This matrix is not Hermitian because $(iT_1)^\dagger = -iT_1 \neq iT_1$. Therefore it cannot be an element of $\mathfrak{su}(2)$.

Example 7.4. Another important group is the Lie group SU(3). As we shall see in Section 7.5, SU(3) is the gauge group of the strong interactions. Since $\dim[\mathfrak{su}(3)] = \dim[\mathrm{SU}(3)] = 3^2 - 1 = 8$, the algebra $\mathfrak{su}(3)$ has eight generators T_A, $A = 1, 2, ..., 8$. It is conventional to define the generators of SU(3) in terms of the eight Gell-Mann matrices λ

$$\lambda_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \lambda_2 = \begin{bmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \lambda_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\lambda_4 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \qquad \lambda_5 = \begin{bmatrix} 0 & 0 & i \\ 0 & 0 & 0 \\ -i & 0 & 0 \end{bmatrix} \qquad \lambda_6 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{7.4}$$

$$\lambda_7 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{bmatrix} \qquad \lambda_8 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix}.$$

As in the case of SU(2), we set $T_A := \frac{1}{2}\lambda_A$, where we adopted the standard convention that capital letters like *A, B, C* can take on every value from 1 to 8.

Let $\mathfrak{g}$ be the Lie algebra of a Lie group $G \subseteq U(n)$ and let T_j $(j = 1, 2, ..., m)$ be generators of $\mathfrak{g}$ (i.e. a basis for $\mathfrak{g}$ as a vector space), where $m = \dim(\mathfrak{g})$. Then for each j, k, $[T_j, T_k]$ can be written uniquely in the form

$$[T_j, T_k] = if_{jkl}T_l := i\sum_{l=1}^{m} f_{jkl}T_l, \tag{7.5}$$

where $[T_j, T_k] := T_jT_k - T_kT_j$ is the commutator of T_j and T_k. The constants f_{jkl} are called the *structure constants* of $\mathfrak{g}$ ([267]). For example, the commutator of the Lie algebra $\mathfrak{su}(2)$ is given by

$$[T_j, T_k] = if_{jkl}T_l := i\sum_{l=1}^{3} \epsilon_{jkl}T_l = i\epsilon_{jkl}T_l, \tag{7.5'}$$

where $T_j = \frac{1}{2}\sigma_i$ $(j = 1, 2, 3)$ and ϵ_{jkl} is the Levi-Civita symbol:

$$\epsilon_{jkl} := \begin{cases} 1 & \text{if } (j,k,l) \in \{(1,2,3),(2,3,1),(3,1,2)\}, \\ 0 & \text{if } j = k \text{ or } k = l \text{ or } l = i, \\ -1 & \text{if } (j,k,l) \in \{(1,3,2),(3,2,1),(2,1,3)\}. \end{cases}$$

Clearly, the structure constants determine the bracket operation on $\mathfrak{g}$. If follows from (7.5) that the commutator of any two generators of a Lie algebra is a linear combination of its generators. Since the generators are Hermitian, the structure constants are real.

The product $[\cdot, \cdot]: \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ is completely determined by the multiplication of basis elements (i.e. generators) of $\mathfrak{g}$. Conversely, once a basis for an algebra $\mathfrak{g}$ has been chosen, the products of basis elements can be set arbitrarily, and then extended in a unique way to a bilinear operator on $\mathfrak{g}$, i.e., the resulting multiplication satisfies the algebra laws. Thus the commutation relations (7.5) (i.e. the structure constants f_{jkl}) define the algebra $\mathfrak{g}$ and (7.5) is simply called the Lie algebra of the group G.

The abstract considerations of Lie groups as above may seem at a first glance like just rather mathematical and formal. Yet, the Lie group theory is of incredible importance in physics. In

[267] It is important to emphasize that the actual value of f_{jkl} is basis dependent. Thus the term 'structure constants' is actually a misnomer.

physical applications, however, one is more interested in what the group actually does, i.e. the group action ([268]). An important idea is that one group can act on many different kinds of objects (this will make much more sense in a moment). This idea motivates the notion of a group representation. ([S4])

Representations allow us to attach a physical meaning to a group element: before introducing the concept of representation, a group element g is just an abstract mathematical object, defined by its composition rules with the other group members. Choosing a specific representation instead allows us to interpret g as a (linear) transformation on a certain space. For example, let V be a real 3-dimensional vector space with an inner product and consider the group SO(3) of all linear transformations L:V → V that preserve the inner product. Now taking as V the set of spatial vectors $\vec{x}$, an abstract element g ∈ SO(3) can be interpreted physically as a rotation R_{g} in physical three-dimensional space. ([M1])

The use of representation theory to exploit the symmetries of a problem has become a powerful tool that has found applications in many areas of science, not just quantum mechanics.

But what exactly is a group representation? In mathematical terms, a representation is defined as follows: let G be a matrix Lie group. A complex [real] *representation* (R, G, V) of G is a group homomorphism ([269])

$$R{:}\mathrm{G} \to \mathrm{GL(V)}, \tag{7.6}$$

where V is a finite dimensional complex [real] vector space. From now on we will assume that a representation is complex unless otherwise specified.

Since V is finite dimensional, say $\mathrm{V} = \mathbb{C}^n$, we can choose a basis of V and identify GL(V) with GL(n). Thus a group representation (7.6) is just a set of n×n matrices, one for each group element, satisfying the multiplication rules of the group: $R(\mathrm{gh}) = R(\mathrm{g})R(\mathrm{h})$ for all g, h ∈ G. The dimension of V is called the *dimension of the representation*.

One should think of a representation as a (linear) action of a group on a vector space, since to every g∈G there is associated a linear operator $R_{\mathrm{g}} := R(\mathrm{g}){:}\mathrm{V} \to \mathrm{V}$, which *acts* on the vector space V ([H0]). We also say that G *acts on* V. In other words, a representation identifies with each element of the abstract group G a linear transformation of a vector space. In this manner, G is 'represented' as a set of linear operators acting on V. The framework of representation theory enables one to examine the group action on very different vector spaces.

A common source of confusion is that representations (R, G, V) are sometimes referred to by the map R, leaving implicit the vector space V that the matrices $R(\mathrm{g})$ act on, but at other times referred to by specifying the vector space V, leaving implicit the map R. ([W14])

In physics, the elements of V are interpreted as values of fields $(\mathrm{t}, \boldsymbol{x}) \to \psi(\mathrm{t}, \boldsymbol{x}) \in \mathrm{V}$ and representation theory provides the appropriate mathematical formalism to describe how symmetries act on them. In particular, finding all representations is then equivalent to finding all fields on which the symmetry can act. These fields in turn can be used to construct any field theory that respects the symmetry.

Each operator R_{g} can be thought as an n×n matrix $\left[(R_{\mathrm{g}})_{ij}\right]$ acting on V via matrix-vector product

$$R_{\mathrm{g}}\mathrm{v} := \left[(R_{\mathrm{g}})_{ij}\right]\mathrm{v}^T = \left(\sum_{j=1}^{n}(R_{\mathrm{g}})_{1j}v_j, \dots, \sum_{j=1}^{n}\left(R_{\mathrm{g}}\right)_{nj}v_j\right) \in \mathrm{V},$$

---

[268] To a physicist, the vector space the group acts upon is called the representation. This is often a space of fields or coordinates.

[269] Given two groups, G and H, a *group homomorphism* from G to H is a function f:G → H such that for all u and v in G it holds that f(uv) = f(u)f(v), so that the function f preserves the group structure. A group homomorphism is called a *group isomorphism* if it is one-to-one and onto.

where v∈V is a vector $v = (v_j)$. Notice that the components of the vector v need not be simply numbers but can be, for example, Dirac spinors.

The fundamental relationship between quantum mechanics and representation theory is that whenever we have a physical quantum system then representation theory provides information about quantum states when G acts on the system ([W14]).

The most interesting classes of representations are those for which the transformations $R(g)$ are *unitary*, i.e. if R:G → U(n) ⊆ GL(V). For a unitary representation (R, G, V), we can use (7.3) to write $R(g)$ as $R(g) = e^{iH}$, where H is a Hermitian matrix. We thus see that a unitary representation (R, G, $\mathcal{H}$) of G acting on a Hilbert space $\mathcal{H}$ of quantum states (see Section 5.9) gives us not just unitary matrices $R(g)$, but also corresponding Hermitian operators H on $\mathcal{H}$. Consequently, Lie group actions provide us with a class of quantum mechanical observables related to these operators. ([W14])

Since a matrix Lie group G is a subgroup of GL(V), it may represent itself, with the homomorphism R being the identity map $g \to R(g) = g \in G \subseteq GL(V)$. Then elements of G acts on V since they are elements of GL(V). In physics, this representation is called the *fundamental representation*. In mathematics, it is called the *defining* or *standard representation*. Accordingly, G can act on very different vector spaces via fundamental representation.

Another example of representation of G is the *trivial representation* (R, G, V) with R being the constant map $g \to R(g) = id_V$, where id_V is the identity on V: $id_V(v) = v$ for v∈V.

It is important to note that the dimension of a representation (R, G, V) is not the same as the dimension of the Lie group G. The dimension of the representation is, by definition, equal to the dimension of the vector space V the group acts on, while the dimension of G is the number of real parameters needed to specify an element of the group. For example, the fundamental representation (R, G, $\mathbb{C}^2$) of SU(2) is two-dimensional because $\mathbb{C}^2$ is two-dimensional. On the other hand, $\dim[SU(2)] = 2^2 - 1 = 3$.

Different representations of a given Lie group need not have the same dimension because they can act on linear spaces of any dimension. So one has to be careful when discussing the dimensions of G and V, as there is a risk of confusion when both are considered simultaneously. ([H17])

Example 7.5. Representations ([270]) of SU(2) are classified by a non-negative integer n = 0, 1, 2, ..., and have dimension n+1. It is common in physics to label these representations by $s = \frac{n}{2} = 0, \frac{1}{2}, 1, \ldots$ and call the representation labelled by s the *spin s representation*.

The lowest dimensional representation is spin 0 representation called *scalar representation.* The objects (scalars) the group acts on in this representation are used to describe elementary particles of spin 0.

The next higher-dimensional representation is called *spin* $\frac{1}{2}$ or *spinor representation* is just the fundamental representation on the space $\mathbb{C}^2$ of spinors, i.e., complex vectors with two components. The objects (spinors) the group acts on in this representation are used to describe elementary particles of spin-½ (i.e. fermions). The significance of the group SU(2) is that every spinor rotation U($\boldsymbol{n}$, θ) defined (see Example 5.3) by

$$U(\boldsymbol{n}, \theta) = e^{i\theta \boldsymbol{n}\cdot\boldsymbol{\sigma}/2}$$

for some axis $\boldsymbol{n}$ and some angle θ is a member of SU(2). Thus SU(2) is the group of rotations on spin-½ systems.

---

[270] More precisely, the irreducible representations of SU(2). An *irreducible* representation is a representation (R, G, V) that has no proper nontrivial subrepresentation. A representation (R, G, W) is a *subrepresentation* of (R, G, V) if W is a vector subspace of V and $R_g(w) \in W$ for all g∈G and w∈W – see e.g. [W14] for details.

The third representation is called *spin* 1 or *vector representation*, because the objects (vectors) the group acts on in this representation are used to describe elementary particles of spin-1 (i.e. bosons). It maps elements of SU(2) into GL($\mathbb{C}^3$). It is actually the adjoint representation ad of the Lie algebra $\mathfrak{su}(2)$ – see Example 7.6. However, representations of $\mathfrak{su}(2)$ are in one-to-one correspondence with representations of SU(2). ([271])

One reason that SU(2) representations are especially tractable is that there is a simple explicit construction of the irreducible representations. Consider the space V_m of homogeneous polynomials in two complex variables with total degree m $\geq$ 1. That is, V_m is the space of functions of the form

$$P(z_1, z_2) = a_0 z_1^m + a_1 z_1^{m-1} z_2 + a_2 z_1^{m-2} z_2^2 + \ldots + a_m z_2^m$$

with $z_1, z_2 \in \mathbb{C}$ and the a_i's arbitrary complex constants. The space V_m is an (m+1)-dimensional complex vector space. If $m = 2$ then the space V_2 is 3-dimensional and consists of functions of the form

$$P(z_1, z_2) = a_0 z_1^2 + a_1 z_1 z_2 + a_2 z_2^2.$$

Let z denote the vector $z = (z_1, z_2) \in \mathbb{C}^2$ and for $U \in SU(2)$ let $U^{-1} = [U_{ij}]$ be the inverse of U. Now define a linear transformation R_U:$V_2 \to V_2$ by the formula

$$\begin{aligned}
&[R_U(P)](z) := P(U^{-1} z^T) = P(U_{11} z_1 + U_{12} z_2, U_{21} z_1 + U_{22} z_2) = \\
&a_0 (U_{11} z_1 + U_{12} z_2)^2 + a_1 (U_{11} z_1 + U_{12} z_2)(U_{21} z_1 + U_{22} z_2) + a_2 (U_{21} z_1 + U_{22} z_2)^2 = \\
&(a_0 U_{11}^2 + a_1 U_{11} U_{21} + a_2 U_{21}^2) z_1^2 + \\
&(2 a_0 U_{11} U_{12} + a_1 U_{11} U_{22} + a_1 U_{12} U_{21} + 2 a_2 U_{21} U_{22}) z_1 z_2 + \\
&(a_0 U_{12}^2 + a_1 U_{12} U_{22} + a_2 U_{22}^2) z_2^2 = \\
&b_0 z_1^2 + b_1 z_1 z_2 + b_2 z_2^2,
\end{aligned}$$

where b_0, b_1, b_2 are complex constants. Thus $R_U(P) \in V_2$. Now, compute

$$\begin{aligned}
[(R_U R_W)(P))](z) &= [R_U(R_W(P))](z) \\
&= R_W(P)(U^{-1} z^T) \\
&= P(W^{-1} U^{-1} z^T) \\
&= [R_{UW}(P)](z).
\end{aligned}$$

It implies that the mapping R:SU(2) $\to$ GL(V_2) satisfies $R(UW) = R(U)R(W)$. Thus (R, SU(2), V_2) is a 3-dimensional (complex) representation of SU(2). Notice that this calculation would not yield the desired result if one defined $[R_U(P)](z) = P(Uz^T)$.

Reasoning similar to the above, one can show that for an arbitrary m $\geq$ 1 (R, SU(2), V_m) is an (m+1)-dimensional representation of the 3-dimensional group SU(2).

The crucial observation is that SU(2) is isomorphic to another group, which is typically designated as Spin(3). Spin(3) is a double cover of the rotation group SO(3), meaning that it is a larger group where each element of SO(3) is associated with two distinct elements of Spin(3). The mapping κ:Spin(3) $\to$ SO(3) which projects back onto SO(3) is a group homomorphism, which is referred to as a covering map. If we have a representation (R, V) of SO(3), then the composition $R \circ \kappa$ gives us a representation of Spin(3) $\cong$ SU(2). Thus, the classification of representations of SU(2) can be considered a classification of representations of SO(3) as well. To be more precise, representations of SU(2) of integer spin are also representations of SO(3). However, representations of SU(2) with half-integer spin are not representations of SO(3).

Generally, we can look at representations of a given group on any vector space. But there is exactly one distinguished vector space that comes automatically with each group G: its own Lie

[271] This is true for every simply connected matrix Lie group – see e.g. [H0].

algebra $\mathfrak{g}$ that is nothing but a real vector space equipped with a Lie bracket. This representation is called the *adjoint representation* ([S5]).

Let $G \subseteq U(n)$ be a matrix Lie group with Lie algebra $\mathfrak{g}$. First, define for all $g \in G$ the *conjugation* (linear) mapping $Ad_g: \mathfrak{g} \to \mathfrak{g}$ that sends $X \in \mathfrak{g}$ to $gXg^{-1} \in \mathfrak{g}$ ([272])

$$X \to Ad_g(X) := gXg^{-1}, \tag{7.7}$$

where gXg^{-1} is the product of matrices. Since g is a unitary matrix, we have $g^{-1} = g^\dagger$ and hence $Ad_g(X) = gXg^\dagger$. Moreover

$$e^{Ad_g(X)} = e^{gXg^{-1}} = ge^X g^{-1}. \tag{7.8}$$

The *adjoint representation* of G is defined as $(Ad, G, \mathfrak{g})$, where

$$Ad : G \to GL(\mathfrak{g}) \text{ ([273])} \tag{7.9}$$

is given by $Ad(g) := Ad_g$. Since $Ad_{gh} = Ad_g Ad_h$, the mapping Ad is a Lie group homomorphism.

Similarly to a group representation, one defines a *representation* $(r, \mathfrak{g}, V)$ of a Lie algebra $\mathfrak{g}$ as a Lie algebra homomorphism ([274]) $r: \mathfrak{g} \to \mathfrak{gl}(V)$, where V is a finite dimensional vector space and $\mathfrak{gl}(V)$ is the Lie algebra of GL(V). And in this case there is also a natural representation of $\mathfrak{g}$ on itself: the *adjoint representation* $(ad, \mathfrak{g}, \mathfrak{gl}(\mathfrak{g}))$, where the homomorphism

$$ad: \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g}), \tag{7.10}$$

is given by

$$\mathfrak{g} \ni Y \to ad(Y) = ad_Y \in \mathfrak{gl}(\mathfrak{g})$$

where ad_Y is defined as the linear map from $\mathfrak{g}$ to itself given by

$$\mathfrak{g} \ni X \to ad_Y(X) := [X, Y]. \tag{7.11}$$

Both adjoint representations Ad and ad are closely related to each other by the exponential function

$$Ad\left(e^Y\right) = e^{ad_Y}. \tag{7.12}$$

Since every action $ad_Y: \mathfrak{g} \to \mathfrak{g}$ is a linear map, so we can represent it using matrices. Let T_j $(j = 1, 2, ..., \dim(\mathfrak{g}))$ be generators of $\mathfrak{g}$ (i.e. a basis for $\mathfrak{g}$ as a vector space). Then the matrix $ad_j = [(ad_j)_{kl}]$, corresponding to ad_{T_j}, can be explicitly constructed from the structure constants (7.5) by defining the matrix elements as

$$(ad_j)_{kl} := -if_{jkl}. \tag{7.13}$$

By this, we mean that the element at row number k and column number l of the matrix ad_j representing the map ad_{T_j} is given by the number $-if_{jkl}$ $(j, k, l = 1, 2, ..., \dim(\mathfrak{g}))$. Hence the structure constants always define the adjoint representation ad. Notice that ad_j is a $\dim(\mathfrak{g}) \times \dim(\mathfrak{g})$ matrix.

Example 7.6. Since the algebra $\mathfrak{su}(2)$ is 3-dimensional, the adjoint representation ad: $\mathfrak{su}(2) \to \mathfrak{gl}(\mathfrak{su}(2))$ can be represented by three 3×3 matrices. Let us compute them explicitly by calculating how basis elements of the (real) linear space $\mathfrak{su}(2)$ act on $\mathfrak{su}(2)$. Taking our basis to be $T_j = \frac{1}{2}\sigma_j$ for $j = 1, 2, 3$ (see Example 7.3), we can write an element $H \in \mathfrak{su}(2)$ as

$$H = \alpha_1 T_1 + \alpha_2 T_2 + \alpha_3 T_3,$$

---

[272] See e.g. [W14] for a proof that gXg^{-1} belongs to $\mathfrak{g}$.

[273] Reminder: $GL(\mathfrak{g})$ denotes the Lie group of all bijective (i.e. one-to-one and onto) linear maps on the linear space $\mathfrak{g}$.

[274] A linear mapping $r: \mathfrak{g} \to \mathfrak{gl}(V)$ is a Lie algebra homomorphism if it maps the Lie brackets on $\mathfrak{g}$ to the Lie brackets on $\mathfrak{gl}(V)$: $r([x, y]) = [r(x), r(y)] := r(x) \circ r(y) - r(y) \circ r(x)$, where $\circ$ is composition of linear maps.

where $\alpha_j \in \mathbb{R}$ ($j = 1, 2, 3$). The adjoint action on H by T_j is the map $ad_{T_j}(\mathrm{H}) = [T_j, \mathrm{H}]$ and it takes the vector $\boldsymbol{\alpha} = [\alpha_1 \quad \alpha_2 \quad \alpha_3]$ to a new vector $\boldsymbol{\beta} = [\beta_1 \quad \beta_2 \quad \beta_3]$, where $\boldsymbol{\beta}^{\mathrm{T}} = J_j \boldsymbol{\alpha}^{\mathrm{T}}$ for some 3x3 matrix J_j. Since the commutator $[\cdot, \cdot]$ is bilinear, we can apply (7.5') to compute $ad_{T_1}(\mathrm{H})$ as

$$\begin{aligned} ad_{T_1}(\mathrm{H}) = [T_1, \mathrm{H}] &= [T_1, \alpha_1 T_1 + \alpha_2 T_2 + \alpha_3 T_3] \\ &= \alpha_1 [T_1, T_1] + \alpha_2 [T_1, T_2] + \alpha_3 [T_1, T_3] \\ &= 0 + \mathrm{i}\alpha_2 T_3 - \mathrm{i}\alpha_3 T_2. \end{aligned}$$

Thus $\boldsymbol{\beta} = [0 \quad -\mathrm{i}\alpha_3 \quad \mathrm{i}\alpha_2]$ and therefore the matrix corresponding to ad_{T_1} is

$$J_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\mathrm{i} \\ 0 & \mathrm{i} & 0 \end{bmatrix}$$

in this basis. The matrices for ad_{T_2} and ad_{T_3} can be obtained similarly

$$J_2 = \begin{bmatrix} 0 & 0 & \mathrm{i} \\ 0 & 0 & 0 \\ -\mathrm{i} & 0 & 0 \end{bmatrix}, \quad J_3 = \begin{bmatrix} 0 & -\mathrm{i} & 0 \\ \mathrm{i} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Notice that the kl entry of the matrix J_j reads $(J_j)_{kl} = -\mathrm{i}\epsilon_{jkl}$. Now, for any $\mathrm{Y} = \kappa_1 T_1 + \kappa_2 T_2 + \kappa_3 T_3 \in \mathfrak{su}(2)$, the adjoint action on $\mathrm{H} \in \mathfrak{su}(2)$ by Y is the map $ad_{\mathrm{Y}}(\mathrm{H}) = [\mathrm{Y}, \mathrm{H}] = \kappa_1 ad_{T_1}(\mathrm{H}) + \kappa_2 ad_{T_2}(\mathrm{H}) + \kappa_3 ad_{T_3}(\mathrm{H})$. Thus the map ad_{Y} is represented by a linear combination of the matrices J_1, J_2, J_3.

Notice that the adjoint representation maps $T_j \to J_j \in \mathrm{GL}(\mathbb{C}^3)$ for $j = 1, 2, 3$. Thus it can be used to define the spin 1 representation of SU(2) which map elements of SU(2) into $\mathrm{GL}(\mathbb{C}^3)$ – see Example 7.5.

The reason why representations are important in the particle physics is because particles can be labelled in terms of how they transform under the group of symmetries, i.e. in terms of the corresponding representations.

There are two sources of symmetries in the Standard Model. One is the Lorentz group (or Poincare group) corresponding to symmetries of spacetime. So we can label particles according to how they transform under the Lorentz group. Another source of symmetries is the gauge group of the Standard Model SU(3)×SU(2)×U(1). The SU(2)×U(1) factor comes from the electroweak force (see Section 7.3), while the SU(3) factor corresponds to the strong force (see Section 7.4). Therefore, particles are labelled by how they transform under this gauge groups: in other words, they are labelled by representations of SU(3), SU(2) and U(1).

This is a general feature that gauge particles (i.e. bosons) transform according to the adjoint representations, whereas the matter particles (i.e. fermions) come in the fundamental representations. ([B5])

One confusing thing is that the term *adjoint representation* is ambiguously used in literature. It is not always clearly stated whether Ad or ad is meant, and it is to be deduced from the context. Notice the difference: the mapping Ad, which takes an element $\mathrm{g} \in \mathrm{G}$ and maps it to Ad_{g}, is a representation of G acting on $\mathfrak{g}$. The mapping ad, on the other hand, which takes an element $\mathrm{Y} \in \mathfrak{g}$ and maps it to ad_{Y}, is a representation of $\mathfrak{g}$ acting on itself. Thus among all these objects of linear mappings one needs to be careful when reading a physics textbook.

To summarise:

- There are always only $\mathrm{m} = \dim(\mathrm{G}) = \dim(\mathfrak{g})$ generators T_j for a group G (i.e. for the algebra $\mathfrak{g}$). They belong to $\mathfrak{g}$ and give elements of G via exponentiation. Thus generators are n×n matrices if and only if the elements of G are n×n matrices.
- $\dim[\mathrm{U}(\mathrm{n})] = \dim[\mathfrak{u}(\mathrm{n})] = \mathrm{n}^2$ and $\dim[\mathrm{SU}(\mathrm{n})] = \dim[\mathfrak{su}(\mathrm{n})] = \mathrm{n}^2 - 1$.

- A group G can act on a linear space V of any dimension, for example via the trivial representation.
- In the adjoint representation Ad of G, the group G acts on the linear space $\mathfrak{g}$.
- In the adjoint representation ad of $\mathfrak{g}$, the algebra $\mathfrak{g}$ acts on itself. The actions ad_Y are defined by the structure constants (7.5) as m×m matrices, where m = dim ($\mathfrak{g}$) = dim(G).
- When we define representations of groups, we map group elements to linear operators in GL(V), that is, invertible matrices. Invertibility is required because of the group properties, since we map the group operation to matrix multiplication. For Lie algebras, we map elements of the vector space to linear operators in $\mathfrak{gl}$(V), i.e. matrices that are not necessarily invertible. This is an important distinction. ([B5])

7.2. Generic construction of gauge theories

We are now in a position to discuss how Lie groups are used to construct a G-gauge theory ([275]) for an arbitrary compact Lie group G ⊆ U(n) ([276]). Its Lie algebra $\mathfrak{g}$ is generated by Hermitian generators (n×n complex matrices) T_j, $j = 1, 2, ..., m$, where m = dim(G) = dim($\mathfrak{g}$). We shall assume that G satisfies (7.3), i.e. every element U ∈ G is of the form

$$U = e^{i\sum_{j=1}^{m}\alpha_j T_j}.$$

Let ψ be a 'matter' field described by a multiplet with n components ([277])

$$(t, \boldsymbol{x}) \to \psi(t, \boldsymbol{x}) = \begin{bmatrix} \psi_1 \\ \psi_2 \\ \cdots \\ \psi_n \end{bmatrix} \tag{7.14}$$

with Lagrangian density

$$\mathcal{L} = \nabla_\mu \psi^\dagger \nabla_\mu \psi - m^2 \psi^\dagger \psi - \lambda(\psi^\dagger \psi)^2 := \sum_{i=1}^{n} [\nabla_\mu \psi_i^* \nabla_\mu \psi_i - m^2 \psi_i^* \psi_i - \lambda(\psi_i^* \psi_i)^2],$$

where $\psi^\dagger = [\psi_1^*, \psi_2^*, ..., \psi_n^*]$ and λ is some function of the square of the modulus of the field. Differentiation of the column means, as usual, differentiation of each of its components (the same for rows or matrices). The column of fields (7.14) can be conceived of as a single field with values in the n-dimensional complex space of columns. It is assumed that the mass of each of the fields ψ_i, $i = 1,2, \ldots, n$, is identical and equal to m. ([R3])

For example, ψ can be the Schrödinger wave function of a spin-0 particle $(t, \boldsymbol{x}) \to \psi(t,\boldsymbol{x}) \in \mathbb{C}$ or the isospin doublet wave function $\psi^{(1/2)}$ considered in Section 6.1.

Every element U of the group G defines the global transformation $\psi' = U\psi$ of the field ψ. Since U is an n×n matrix we can write this transformation as

$$\psi' = U\psi = [u_{kl}]\psi, \tag{7.15}$$

where $U = [u_{kl}]$ and $\psi'_k = u_{kl}\psi_l := \sum_{l=1}^{n} u_{kl}\psi_l$. Thus the transformation (7.15) mixes the components ψ_k inside the multiplet ψ. Notice that (7.15) is a transformation under the action of

---

[275] A *G-gauge theory* is a field theory that has a gauge symmetry induced by a gauge group of transformations G, which is required to be a Lie group. The field equations of a G-gauge theory are covariant with respect to the group G.

[276] Since U(n) is compact, it is enough to assume that G is closed in U(n), for closed subsets of a compact set are compact. We assume that G is compact because then it has a unitary, finite dimensional and irreducible representation. We will use this property only implicitly. A detailed discussion of irreducible representations (see e.g. [S3]) would lead us too far astray here.

[277] Notice that the components of the wave function ψ need not be simply complex numbers but can be something more intricate such as, for example, Dirac spinors: $\psi_i(t, \boldsymbol{x}) \in \mathbb{C}^4$, $i = 1, 2, ..., n$.

the fundamental representation of the group G. Indeed, according to (7.15), the matrix $R(\mathrm{U}) = \mathrm{U}$ acts on the field ψ. ([278])

If we wish to promote the global symmetry to a local one by letting U become a function of spacetime events $(\mathrm{t}, \boldsymbol{x})$, it is convenient to write U in exponential form

$$\mathrm{U}(\mathrm{t}, \boldsymbol{x}) = e^{\mathrm{i}\sum_{j=1}^{m} \alpha_j(t,\boldsymbol{x})T_j}. \tag{7.16}$$

Then (7.15) amounts to

$$\psi(\mathrm{t}, \boldsymbol{x}) \to \psi'(\mathrm{t}, \boldsymbol{x}) = \mathrm{U}(\mathrm{t}, \boldsymbol{x})\psi(\mathrm{t}, \boldsymbol{x}) = e^{\mathrm{i}\sum_{j=1}^{m} \alpha_j(t,\boldsymbol{x})T_j}\psi(\mathrm{t}, \boldsymbol{x}) = e^{\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{T}}\psi(\mathrm{t}, \boldsymbol{x}), \tag{7.17}$$

where $\boldsymbol{T} := (T_1, T_2, ..., T_m)$ are generators of $\mathfrak{g}$ and $\boldsymbol{\alpha} := (\alpha_1, \alpha_2, ..., \alpha_m)$ are real functions $(\mathrm{t}, \boldsymbol{x}) \to \alpha_j(\mathrm{t}, \boldsymbol{x})$, $j = 1, 2, ..., \mathrm{m}$.

The Lagrangian density of a Yang-Mills theory should be unchanged when a vector $\psi(\mathrm{t}, \boldsymbol{x})$ transforms to $\psi'(\mathrm{t}, \boldsymbol{x}) = e^{\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{T}}\psi(\mathrm{t}, \boldsymbol{x})$. This mapping has the 'generalized phase transformation' character of (6.9), now with m 'phase angles' $\boldsymbol{\alpha}$.

The gauge covariance requires that the field equations for the transformed fields must have the same form as the original ones. Since the field equations always contain derivatives of the fields, the transformation (7.17) leads to additional m terms that break the covariance due to the $(\mathrm{t}, \boldsymbol{x})$-dependence of the parameters $\alpha_j(\mathrm{t}, \boldsymbol{x})$. Indeed, under the transformation (7.17), the derivative of the field becomes (by the usual product rule)

$$\nabla_\mu\psi'(\mathrm{t}, \boldsymbol{x}) = \nabla_\mu[\mathrm{U}(\mathrm{t}, \boldsymbol{x})\psi(\mathrm{t}, \boldsymbol{x})] = \mathrm{U}(\mathrm{t}, \boldsymbol{x})\nabla_\mu\psi(\mathrm{t}, \boldsymbol{x}) + [\nabla_\mu\mathrm{U}(\mathrm{t}, \boldsymbol{x})]\psi(\mathrm{t}, \boldsymbol{x})$$

i.e. $\nabla_\mu\psi$ does not transform in the same manner as ψ. The derivative $\nabla_\mu\psi'$ contains now the term

$$\nabla_\mu\mathrm{U}(\mathrm{t}, \boldsymbol{x}) = \nabla_\mu e^{\mathrm{i}\sum_{j=1}^{m} \alpha_j(t,\boldsymbol{x})T_j} = \mathrm{iU}(\mathrm{t}, \boldsymbol{x})\sum_{j=1}^{m}(\nabla_\mu\alpha_j(t, \boldsymbol{x}))T_j,$$

which includes m derivatives $\nabla_\mu\alpha_j(\mathrm{t}, \boldsymbol{x})$, one for each generator T_j. Consequently, in contrast to the U(1) case, in order to make the Lagrangian locally G invariant, it is no longer sufficient to introduce one single compensating field. Instead, a set of m fields is required to cancel the unwanted term $(\nabla_\mu\mathrm{U})\psi$.

To restore the covariance the field equations we replace the conventional derivative ∇_μ with the covariant derivative D_μ and require that under the transformations (7.17) it becomes $(D_\mu\psi)' = \mathrm{U}D_\mu\psi$. This can be accomplished by introducing a gauge field $W_\mu(\mathrm{t}, \boldsymbol{x})$ and writing $D_\mu\psi$ in the form $D_\mu\psi = \nabla_\mu\psi - \mathrm{ig}W_\mu\psi$, where the coupling strength g is added for later convenience.

The next step is to determine the structure of the field W_μ (referred to as the *Yang–Mills field)* by defining the range of its values. This may be achieved by examining the manner in which the field W_μ must behave under gauge transformations (7.16). It turns out (see e.g. [R3]) that $(\mathrm{t}, \boldsymbol{x}) \to W_\mu(\mathrm{t}, \boldsymbol{x})$ is a field with values in the Lie algebra $\mathfrak{g}$ of the group G. More precisely, the component functions $W_0(\mathrm{t}, \boldsymbol{x})$, $W_1(\mathrm{t}, \boldsymbol{x})$, $W_2(\mathrm{t}, \boldsymbol{x})$ and $W_3(\mathrm{t}, \boldsymbol{x})$ take values in the Lie algebra $\mathfrak{g}$. ([279])

Thus, $W_\mu(\mathrm{t}, \boldsymbol{x})$ is a Hermitian n×n complex matrix $[w_{kl}^{\mu}(\mathrm{t}, \boldsymbol{x})]$, $k, l = 1, 2, ..., n$, for each $(\mathrm{t}, \boldsymbol{x})$ and $\mu = 0, 1, 2, 3$. An alternative perspective on the gauge field $(\mathrm{t}, \boldsymbol{x}) \to W_\mu(\mathrm{t}, \boldsymbol{x})$ is to regard it as

---

[278] Here R acts on the vector space of quantum mechanical wave functions. Although this vector space is infinite dimensional, essentially in every case of physical interest we can take a properly chosen finite dimensional vector subspace. Consequently, we can apply finite dimensional representations also to this case. We can think of the vector space V, R acts on, as the n-dimensional complex space of columns (7.14). Then ψ can be understood as a field $(\mathrm{t}, \boldsymbol{x}) \to \psi(\mathrm{t}, \boldsymbol{x}) \in \mathrm{V}$, i.e. ψ takes values in the space of the representation (R, G, GL(V)).

[279] Recall that the same symbol is used for the vector W_μ and its components W_μ, $\mu = 0, 1, 2, 3$.

matrix-valued, where $W_\mu(\mathrm{t}, \boldsymbol{x})$ is an n×n matrix whose elements are vector fields $(\mathrm{t}, \boldsymbol{x}) \to w^\mu_{kl}(\mathrm{t}, \boldsymbol{x}) = [w^0_{kl}(\mathrm{t}, \boldsymbol{x}), w^1_{kl}(\mathrm{t}, \boldsymbol{x}), w^2_{kl}(\mathrm{t}, \boldsymbol{x}), w^3_{kl}(\mathrm{t}, \boldsymbol{x})] \in \mathbb{C}^4$.

The elements of the matrix $W_\mu(\mathrm{t}, \boldsymbol{x})$ are still unknown, and to find them will be our next problem. Since each component $(\mathrm{t}, \boldsymbol{x}) \to W_\mu(\mathrm{t}, \boldsymbol{x})$, $\mu = 0, 1, 2, 3$, takes values in the Lie algebra $\mathfrak{g}$, it can be written as a linear combination of the generators T_j, $j = 1, 2, ..., \mathrm{m}$, in the form

$$W_\mu(\mathrm{t}, \boldsymbol{x}) = W^1_\mu(\mathrm{t}, \boldsymbol{x})T_1 + W^2_\mu(\mathrm{t}, \boldsymbol{x})T_2 + ... + W^m_\mu(\mathrm{t}, \boldsymbol{x})T_m, \tag{7.18}$$

where for each value of the indices $\mu = 0, 1, 2, 3$ and $j = 1, 2, ..., \mathrm{m}$, $(\mathrm{t}, \boldsymbol{x}) \to W^j_\mu(\mathrm{t}, \boldsymbol{x})$ is a real-valued function (recall that $\mathfrak{g}$ is a real vector space). These functions build m compensating *gauge vector fields*

$$(\mathrm{t}, \boldsymbol{x}) \to W^j_\mu(\mathrm{t}, \boldsymbol{x}) = [W^j_0(\mathrm{t}, \boldsymbol{x}), W^j_1(\mathrm{t}, \boldsymbol{x}), W^j_2(\mathrm{t}, \boldsymbol{x}), W^j_3(\mathrm{t}, \boldsymbol{x})],$$

one vector field for each generator T_j, $j = 1, 2, ..., \mathrm{m}$.

Thus, the matrix gauge field $(\mathrm{t}, \boldsymbol{x}) \to W_\mu(\mathrm{t}, \boldsymbol{x})$ can be expressed as a linear combination of m vector gauge fields $(\mathrm{t}, \boldsymbol{x}) \to W^j_\mu(\mathrm{t}, \boldsymbol{x})$. It is important to notice that the gauge fields $W^j_\mu(\mathrm{t}, \boldsymbol{x})$ are necessarily real-valued vector fields, and that they arise as 'coordinates' of the more fundamental objects $W_\mu(\mathrm{t}, \boldsymbol{x})$. Selecting a different set of generators of $\mathfrak{g}$ leads to the same W_μ, but to a different set of fields W^j_μ. The vector fields $(\mathrm{t}, \boldsymbol{x}) \to W^j_\mu(\mathrm{t}, \boldsymbol{x})$, $j = 1, 2, ..., \mathrm{m}$, are associated with the physical gauge bosons of the theory. ([W8])

Since each generator T_j belongs to $\mathfrak{g}$, so we can write it as an n×n Hermitian matrix $T_j = [t^j_{kl}]$, $k, l = 1, 2, ..., n$. Consequently the elements of the matrix $W_\mu = [w^\mu_{kl}]$ can be written as

$$w^\mu_{kl} = \sum_{j=1}^m t^j_{kl} W^j_\mu,$$

i.e. the elements w^μ_{kl} of W_μ are linear combinations of the fields W^j_μ ($j = 1, 2, ..., \mathrm{m}$).

Let us denote $\boldsymbol{W}_\mu := (W^1_\mu(\mathrm{t}, \boldsymbol{x}), W^2_\mu(\mathrm{t}, \boldsymbol{x}), ..., W^m_\mu(\mathrm{t}, \boldsymbol{x}))$. Then using (7.18) we may write W_μ as

$$W_\mu = \sum_{j=1}^m T_j W^j_\mu = \boldsymbol{T}\cdot\boldsymbol{W}_\mu. \tag{7.18'}$$

Now applying the gauge principle (cf. Section 5.4) we have to replace the derivative ∇_μ by the covariant derivative D_μ

$$D_\mu := \mathrm{I}_n\nabla_\mu - \mathrm{i}gW_\mu \tag{7.19}$$

where I_n is the n×n identity matrix and g a coupling constant. From the context it should be clear that now D_μ is matrix-valued. The derivative D_μ acts on the n-component field (7.14). The notation $D_\mu\psi$ should be understood as $\left[(D_\mu\psi)_1 \;\; (D_\mu\psi)_2 \;\; ... \;\; (D_\mu\psi)_\mathrm{n}\right]^T$, where $(D_\mu\psi)_\mathrm{k} = \nabla_\mu\psi_k - \mathrm{i}g\sum_{l=1}^n w^\mu_{kl}\psi_l$ ([280]). The parameter g determines the coupling strength between the 'matter' field ψ and the gauge field W_μ.

The key property required for the covariant derivative D_μ is that under the transformation (7.15) it becomes

$$(D_\mu\psi)' = \mathrm{U}D_\mu\psi, \tag{7.20}$$

where $\mathrm{U} = e^{\mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{T}}$. It imposes additional constraints on the gauge field W_μ, because (7.20) amounts to

$$(\nabla_\mu - \mathrm{i}gW'_\mu)\mathrm{U}\psi = \mathrm{U}(\nabla_\mu - \mathrm{i}gW_\mu)\psi$$

---

[280] By abuse of notation, $(D_\mu\psi)_\mathrm{k}$ is often written as $D_\mu\psi_k$.

where ∇_μ is multiplied implicitly by the identity matrix I_n. It turns out that the required transformation for W_μ is ([R3])

$$W'_\mu = UW_\mu U^\dagger - \frac{i}{g}(\nabla_\mu U)U^\dagger. \quad (7.21)$$

Then $D_\mu\psi$ transforms in the same way as ψ. In particular, W_μ transforms under a global gauge transformation U as

$$W_\mu \to UW_\mu U^\dagger. \quad (7.22)$$

Since $U^\dagger = U^{-1}$, we infer that W_μ transforms in the adjoint representation Ad of G. We note one of the main differences of the non-Abelian field W_μ from the vector potential of electrodynamics A_μ. Under global transformations A_μ does not change, but non-Abelian potentials transform non-trivially in the adjoint representation in line with (7.22). The reason is that U and $U^\dagger = U^{-1}$ cannot be brought together in (7.22), because they do not commute with the field W_μ.

It is a general rule that in a G-gauge theory, gauge covariance requires that matter fields transform in the fundamental representation, whereas gauge fields are required to transform according to the adjoint representation Ad of G.

The covariant derivative D_μ of the field ψ describes the coupling of the gauge bosons to the particles of the field ψ (i.e. its component fields). In contrast, the commutator $[W_\mu, W_\nu]$ describes the interaction of the gauge fields with themselves (if the group G is non-Abelian).

A physical theory that incorporates the gauge field W_μ must treat W_μ as a dynamical field, and thus the action should contain a kinetic term for W_μ. In other words, the action should include derivative terms for W_μ, which can be found in the field strength.

Therefore, the next step is to identify the strength tensor $F_{\mu\nu}$ for the field W_μ. In analogy with electrodynamics, it is anticipated that the strength tensor will contain a term of the form

$$\nabla_\mu W_\nu - \nabla_\nu W_\mu. \quad (7.23)$$

It is evident from (7.22) that the expression (7.23) must transform according to the adjoint representation of the group in the case of global transformations. We require the strength tensor to transform according to the adjoint representation for all gauge transformations. That is to say, $F_{\mu\nu}$ must be 'gauge covariant'([282])

$$F_{\mu\nu} \to F'_{\mu\nu} = U(t, \boldsymbol{x})F_{\mu\nu}U^\dagger(t, \boldsymbol{x}).$$

However, expression (7.23) itself does not have this property. Indeed, differentiating (7.21), it can be shown ([R3]) that

$$\begin{aligned}\nabla_\mu W'_\nu - \nabla_\nu W'_\mu &= \nabla_\mu[UW_\nu U^\dagger - \frac{i}{g}(\nabla_\nu U)U^\dagger] - \nabla_\nu[UW_\mu U^\dagger - \frac{i}{g}(\nabla_\mu U)U^\dagger] \\ &\neq U(\nabla_\mu W_\nu - \nabla_\nu W_\mu)U^\dagger.\end{aligned}$$

Therefore, we cannot simply define $F_{\mu\nu}$ as the expression (7.23) above. The appropriate generalisation of the electromagnetic field strength tensor $F_{\mu\nu}$ turns out to be

$$\boldsymbol{F}_{\mu\nu} := \nabla_\mu \boldsymbol{W}_\nu - \nabla_\nu \boldsymbol{W}_\mu + g\boldsymbol{W}_\mu \times \boldsymbol{W}_\nu, \quad (7.24)$$

where $\boldsymbol{F}_{\mu\nu} := (F^1_{\mu\nu}, F^2_{\mu\nu}, ..., F^m_{\mu\nu})$,

$$F^j_{\mu\nu} = \nabla_\mu W^j_\nu - \nabla_\nu W^j_\mu + g\sum_{k,l=1}^m f_{jkl}W^k_\mu W^l_\nu, j = 1, 2, ..., m$$

and f_{jkl} are the structure constants of $\mathfrak{g}$ – see (7.5). Writing $F_{\mu\nu} := \boldsymbol{T}\cdot\boldsymbol{F}_{\mu\nu}$ we obtain

---

[281] If G is Abelian, for example if G = U(1), the transformation $W_\mu \to UW_\mu U^\dagger - \frac{i}{q}(\nabla_\mu U)U^\dagger$ amounts to $W_\mu \to W_\mu - \nabla_\mu\Lambda$ (see Section 6.2).

[282] In the non-Abelian case there is no gauge invariant field tensor. Thus the next best possibility is to consider a 'gauge covariant' quantity ([H13]).

$$F_{\mu\nu} = \nabla_\mu W_\nu - \nabla_\nu W_\mu + \mathrm{ig}[W_\mu, W_\nu] = \frac{1}{\mathrm{ig}}[D_\mu, D_\nu]. \ (^{283})$$

At this point, the manner in which we introduce this field strength tensor may appear somewhat opaque. But, of course, there is a deep reason why the correct field strength tensor must be the commutator of the covariant derivative. However, a proper discussion lies beyond the scope of this paper (284). ([S3])

The field strength tensor (7.24) determines the 'kinetic term' of the Lagrangian density of the G-gauge theory

$$\mathcal{L}_G^{kinetic} = -\frac{1}{4}\boldsymbol{F}_{\mu\nu}\boldsymbol{F}^{\mu\nu} := -\frac{1}{4}\sum_j F^j_{\mu\nu} F^{j\,\mu\nu}. \tag{7.25}$$

The remaining terms of the (invariant) Lagrangian density are defined using the covariant derivative (7.19). A major new feature of (7.25) as compared with the Lagrangian density $\mathcal{L}_{em}^{free}$ is that it is not a free Lagrangian, but contains self-interactions, precisely because of the additional term in (7.24) ([J3]).

In the above we have not specified what the label j on $\boldsymbol{W}_\mu$ refers to. It has simply been assumed that a multiplet $\psi(t, \boldsymbol{x})$ is transforming according to some representation of the symmetry group, with j denoting the different components of the field $\boldsymbol{W}_\mu$. In the context of electroweak theory, where the non-Abelian part of the group is SU(2) the W_μ^j will form a triplet, where j essentially labels the weak charge. For QCD where the non-Abelian group is SU(3), j will be a colour-anticolour label (see Example 7.9).

To summarise: we need m vector fields $\boldsymbol{W}_\mu = (W_\mu^1, W_\mu^2, ..., W_\mu^m)$ to cancel the terms that make our Lagrangian non-invariant under local G transformations and must introduce these new fields in such a way that they are able to cancel terms that involve derivatives of $\alpha_j(t, \boldsymbol{x})$. The fields W_μ^j ($j = 1, 2, ..., m$) are associated with the physical gauge bosons of the theory.

The (covariant) field equations (equations of motion) of this gauge theory are then derived from the (invariant) Lagrangian density as Euler-Lagrange equations. The group G is called the *gauge group* (or *structure group*) (285) of the theory, which in turn is called G-*gauge theory*. After the field equations have been formulated, all fields of the theory have to be quantised so that the G-gauge theory becomes a QFT.

Notice that the Lagrangian density of a G-gauge theory must not contain the (non-invariant) mass term $\frac{1}{2}m^2\boldsymbol{W}_\mu\boldsymbol{W}^\mu$. Thus the bosons associated with the gauge fields W_μ^j ($j = 1, 2, ..., m$) must be massless.

As we have seen above, one can construct a gauge theory starting with an arbitrary (compact) Lie group $G \subseteq U(n)$. It turns out, however, that only a few of them are relevant for the Standard Model. These are U(1), SU(2) and SU(3). Let us discuss them briefly. ([H1])

Example 7.7. Let $G = U(1)$ be the (Abelian) group of all unitary one-dimensional matrices. As we have seen in Section 5.5, the group U(1) corresponds to the circle group, consisting of all complex numbers with absolute value 1 under multiplication. It is a 1-dimensional, compact Lie group and the algebra $\mathfrak{u}(1) = \mathbb{R}$ has exactly one generator T, which a (nonzero) 1×1 matrix, i.e.

---

283 The factor $1/\mathrm{ig}$ is needed to cancel the factors that come from the definition (7.19).

284 In mathematical terms, the field strength tensor describes the curvature of a connection on a fibre bundle. The curvature is derived from the connection essentially by taking commutators of certain differential operators related to the connection. In our case, $W_\mu = \boldsymbol{T}\cdot\boldsymbol{W}_\mu$ is a connection on the G principal bundle. This is a rather abstract object, taking values in the Lie algebra $\mathfrak{g}$. For G = SU(n), a more down to earth perspective is to view W_μ simply as a traceless n×n Hermitian matrix. In physical terms, zero curvature indicates 'no physical effect', while $F_{\mu\nu} \neq 0$ implies the presence of a physical effect.

285 Strictly speaking, the gauge group is the much bigger group of maps from spacetime into G because we deal with local symmetry. It is the infinite product $\prod_x G_x$, where G_x is a copy of G for $x = (t, \boldsymbol{x}) \in \mathbb{M}$.

T is just a (real) number (Example 7.1). Consequently, we need only one 'compensating' gauge vector field W_μ. The component fields $W_0(t,\boldsymbol{x})$, $W_1(t,\boldsymbol{x})$, $W_2(t,\boldsymbol{x})$, $W_3(t,\boldsymbol{x})$ are real-valued, i.e. they take values in the Lie algebra of U(1).

The fundamental representation of U(1) is one-dimensional. Thus if $(t, \boldsymbol{x}) \to \psi(t,\boldsymbol{x}) \in \mathbb{C}$ is a complex field (e.g. the Schrödinger wave function of a spin-0 particle) then ψ transforms in the fundamental representation as

$$\psi(t, \boldsymbol{x}) \to \psi'(t, \boldsymbol{x}) = e^{i\alpha(t,\boldsymbol{x})T}\psi(t, \boldsymbol{x}), \tag{7.26}$$

i.e. the fundamental representation yields local phase transformations considered in Section 5.4.

The adjoint representation of U(1) is trivial: Ad_U is the identity mapping for each $U \in U(1)$. Indeed, $Ad_U(X) = e^{i\alpha}Xe^{-i\alpha} = X$, where $U = e^{i\alpha}$. The gauge field W_μ transforms in the adjoint representation according to (7.21)

$$W_\mu \to UW_\mu U^\dagger - \frac{i}{T}(\nabla_\mu U)U^\dagger = W_\mu - \nabla_\mu \alpha, \tag{7.27}$$

where $U(t, \boldsymbol{x}) = e^{iT\alpha(t,\boldsymbol{x})}$ is a local U(1) transformation corresponding to the generator T of $\mathfrak{u}(1)$.

There are two important examples of U(1)-gauge theories. First, let us take the generator $T = q$, where q is electric charge. The resulting U(1) gauge theory is then electromagnetism, denoted $U(1)_{em}$. The gauge transformation (7.17) has now the same form as (5.10)

$$\psi(t, \boldsymbol{x}) \to \psi'(t, \boldsymbol{x}) = U(t, \boldsymbol{x})\psi(t, \boldsymbol{x}) = e^{iq\alpha(t,\boldsymbol{x})}\psi(t, \boldsymbol{x}).$$

The 'compensating' gauge field W_μ can be interpreted as electromagnetic 4-potential A_μ. Moreover, $g = -q$, therefore (7.19) and (7.24) amount to

$$D_\mu := \nabla_\mu - igW_\mu = \nabla_\mu + iqA_\mu,$$

and

$$F_{\mu\nu} = \nabla_\mu A_\nu - \nabla_\nu A_\mu,$$

respectively. Finally

$$\mathcal{L}^{kinetic}_{U(1)_{em}} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} = \mathcal{L}^{free}_{em}.$$

The second important example is the $U(1)_Y$ gauge theory, where Y stands for the *weak hypercharge*, also denoted Y_W. Without getting into details, let us only say that Y is a quantum number relating the electric charge and the third component of weak isospin (see Section 7.4). Now $T = Y$ and (7.17) amounts to

$$\psi(t, \boldsymbol{x}) \to \psi'(t, \boldsymbol{x}) = U(t, \boldsymbol{x})\psi(t, \boldsymbol{x}) = e^{iY\alpha(t,\boldsymbol{x})}\psi(t, \boldsymbol{x}).$$

Again, we need only one 'compensating' gauge field, that we denote B_μ. The covariant derivative is then

$$D_\mu := \nabla_\mu - ig'YB_\mu,$$

where g' is a coupling constant, and the Lagrangian density

$$\mathcal{L}^{kinetic}_{U(1)_Y} = -\frac{1}{4}G_{\mu\nu}G^{\mu\nu},$$

where

$$G_{\mu\nu} = \nabla_\mu B_\nu - \nabla_\nu B_\mu.$$

Example 7.8. Now let $G = SU(2)$ and take the generators $T_j := \frac{1}{2}\sigma_j$, $j = 1, 2, 3$, where σ_j are the three Pauli matrices (see Example 7.3). Let us write the G-gauge transformation (7.17) for SU(2) of a doublet $\psi = [\psi_1 \quad \psi_2]^T$ of two spin-$\frac{1}{2}$ fields (e.g. nucleon isospin doublet $\psi^{(1/2)} = [\psi_p \quad \psi_n]^T$ – see Section 6.1)

(7.28) $$\psi(t, \boldsymbol{x}) \to \psi'(t, \boldsymbol{x}) = e^{i\sum_{j=1}^{3} \alpha_j(t,\boldsymbol{x}) T_j} \psi(t, \boldsymbol{x}).$$

Putting $\boldsymbol{\alpha} := (\alpha_1, \alpha_2, \alpha_3)$ and $\boldsymbol{T} := (T_1, T_2, T_3) = \frac{1}{2}(\sigma_1, \sigma_2, \sigma_3) = \frac{1}{2}\boldsymbol{\tau}$, we can see that (7.28) amounts to the Yang-Mills transformation (6.10)

$$\psi(t, \boldsymbol{x}) \to \psi'(t, \boldsymbol{x}) = e^{i\boldsymbol{\alpha}(t,\boldsymbol{x}) \cdot \boldsymbol{T}} \psi(t, \boldsymbol{x}) = e^{i\frac{1}{2}\boldsymbol{\alpha}(t,\boldsymbol{x}) \cdot \boldsymbol{\tau}} \psi(t, \boldsymbol{x}).$$

In other words, an isospin doublet wave function transforms (locally) in the fundamental representation of the group SU(2).

The Lagrangian density, in the absence of any interactions, is (cf. Eq. (5.29))

(7.29) $$\mathcal{L} := \mathcal{L}^{free}_{Dirac}(\psi_1) + \mathcal{L}^{free}_{Dirac}(\psi_2) = (i\,\overline{\psi}_1\gamma_\mu\nabla^\mu\psi_1 - m_1\overline{\psi}_1\psi_1) + (i\,\overline{\psi}_2\gamma_\mu\nabla^\mu\psi_2 - m_2\overline{\psi}_2\psi_2),$$

where m_i is the mass of the field ψ_i. Notice that ψ_1 and ψ_2 are four-component Dirac spinors (5.27) and the Dirac matrices γ_μ act on these components. Denoting by M the following mass matrix

$$M = \begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix},$$

we can write the Lagrangian density (7.29) for the doublet ψ as

(7.30) $$\mathcal{L} = i\overline{\psi}\gamma_\mu\nabla^\mu\psi - \overline{\psi}M\psi.$$

If the two masses are equal $m_1 = m_2 = m$, the Lagrangian density becomes

$$\mathcal{L} = \overline{\psi}(i\gamma_\mu\nabla^\mu - m)\psi.$$

This looks just like the Dirac Lagrangian density (5.29). However, ψ is now a doublet, i.e. a two-element column vector ([G6]).

This Lagrangian density is not gauge invariant under local SU(2) transformations. In order to guarantee its gauge invariance we have to apply the gauge principle (cf. Section 5.2). First, we have to introduce three vector fields W_μ^j, j = 1, 2, 3, one field for each generator T_j. And then we have to replace the derivative ∇_μ by the covariant derivative D_μ

$$D_\mu := \nabla_\mu - ig\sum_{j=1}^{3} T_j W_\mu^j = \nabla_\mu - ig\frac{1}{2}\boldsymbol{\tau} \cdot \boldsymbol{W}_\mu,$$

where $\boldsymbol{W}_\mu = (W_\mu^1, W_\mu^2, W_\mu^3)$ and g is a coupling strength. The gauge matrix field $W_\mu := \frac{1}{2}\boldsymbol{\tau} \cdot \boldsymbol{W}_\mu$ transforms according to (7.21), i.e. in the adjoint representation of SU(2). The complete SU(2) Lagrangian density becomes

(7.30') $$\mathcal{L}_{SU(2)} = \overline{\psi}(i\gamma_\mu D^\mu - m)\psi - \frac{1}{4}\boldsymbol{F}_{\mu\nu}\boldsymbol{F}^{\mu\nu},$$

where $\boldsymbol{F}_{\mu\nu} := \nabla_\mu \boldsymbol{W}_\nu - \nabla_\nu \boldsymbol{W}_\mu + g\boldsymbol{W}_\mu \times \boldsymbol{W}_\nu$ is the SU(2) field strength tensor (cf. Eq. (7.24)). The Lagrangian density $\mathcal{L}_{SU(2)}$ is invariant under SU(2) gauge transformations. It describes two equal mass Dirac fields interacting with three massless vector gauge fields.

Notice that the SU(2) field strength tensor is equal to Eq. (6.16) and the Lagrangian density (7.30') amounts to the Lagrangian density $\mathcal{L}_{YM}$ (see Eq. (6.19)). Consequently, the original Yang-Mills theory corresponds to a gauge theory constructed with the gauge group SU(2).

The Yang-Mills theory in its original form turned out to be of little use. After all, it assumes that there exist two (distinct) elementary spin-$\frac{1}{2}$ particles of equal mass which have never been observed in nature. However, the Yang-Mills theory is still important because it serves as a prototype of non-Abelian gauge theories, that is, theories for which the generators of the underlying symmetry group do not commute. When non-Abelian gauge theory finally came into its own, it was in the context of colour SU(3) symmetry in the strong interactions ([G6]). Moreover, as we shall see in Section 7.4, the gauge group SU(2) occurs in the GSW model of electroweak interactions. There are, however, some important differences. First, the gauge

group is not simply SU(2), but $SU(2)_L \times U(1)_Y$, so there are four gauge bosons. Second, in a pure gauge theory the gauge bosons are massless. To overcome this difficulty, the GWS model makes use of the phenomenon of spontaneous symmetry breaking. ([J3])

Example 7.9. Quantum chromodynamics (QCD) is a Yang–Mills theory with gauge group SU(3) – cf. Section 7.5. In QCD, the strong interaction is invariant under rotations in colour space i.e. the same for all three colours (r, g, b) = ('read, blue, green') ([286]). This SU(3) symmetry is exact, unlike the approximate uds flavour symmetry.

The Lie algebra $\mathfrak{su}(3)$ has $3^2 - 1 = 8$ generators. They are given by eight 3×3 Gell-Mann matrices (7.4). The gauge transformation (7.17) acts now on quark field ψ, which is an SU(3) triplet $\psi = [\psi_r \quad \psi_b \quad \psi_g]^T$ with colour index. It means that ψ transforms (locally) in the fundamental representation of the group SU(3).

The next step is to introduce eight gauge fields

$$G_\mu^A,\ A = 1, 2, 3, ..., 8,$$

one field for each generator T_A. The covariant derivative D_μ is defined as

$$D_\mu := \nabla_\mu - ig\sum_{A=1}^{8} T_A\, G_\mu^A = \nabla_\mu - igG_\mu,$$

where

$$G_\mu := \boldsymbol{T}\cdot\boldsymbol{G}_\mu = \sum_{A=1}^{8} T_A\, G_\mu^A$$

and $\boldsymbol{G}_\mu := (G_\mu^1, G_\mu^2, ..., G_\mu^8)$. The gauge field G_μ transforms according to (7.21), i.e. in the adjoint representation of SU(3). The derivative transformation in adjoint representation is significant for describing the internal dynamics of the theory. The field strength tensor $\boldsymbol{F}_{\mu\nu}$ is

$$\boldsymbol{F}_{\mu\nu} := \nabla_\mu \boldsymbol{G}_\nu - \nabla_\nu \boldsymbol{G}_\mu + g\boldsymbol{G}_\mu \times \boldsymbol{G}_\nu. \tag{7.31}$$

As we shall see in Section 7.5, the SU(3) gauge theory models the theory of strong interactions. The coupling g is denoted then by g_s whereas ψ describes a colour triplet, i.e. ψ is a three-quark wave function

$$\psi(t, \boldsymbol{x}) = \left[\psi_r(t, \boldsymbol{x}) \quad \psi_b(t, \boldsymbol{x}) \quad \psi_g(t, \boldsymbol{x})\right]^T \in \mathbb{C}^4 \times \mathbb{C}^4 \times \mathbb{C}^4$$

with colour index. The SU(3) local symmetry is not broken and the eight spin-1 gauge fields G_μ^A, called *gluon fields*, are massless.

To summarise the examples above:

Interaction		Gauge group	Number of gauge fields	Field quanta	Charge
electroweak	electromagnetic	SU(2)×U(1)	1	massless photon	electric charge
	weak		3	massive W- and Z-bosons (Higgs needed)	weak isospin
strong		SU(3)	8	massless gluons	colour

7.3. The Brout-Englert-Higgs mechanism ([287])

As we have seen in Section 6.2, Yang and Mills tried to develop a non-Abelian gauge theory to make hadronic isospin into a local SU(2) symmetry. But this formalism turned out not to describe interactions between hadrons. Instead, the local SU(2) symmetry (called *weak isospin*)

[286] This is merely a picturesque way of referring to the three basic states of this degree of freedom and has nothing to do with real colour – see Section 7.5.

[287] See [H9] for historical exposition of this subject and [A1] as a general reference for this section.

describes the weak interactions between the constituents of the hadrons, namely quarks – and leptons. Together with QED, it constitutes the electroweak theory – see Section 7.4.

However, as we know from Chapter 2, the weak interactions are short-ranged, so that their mediating quanta $W^{\pm}$ and Z^{0} must be massive. At first sight, this seems to rule out the possibility of a gauge theory of weak interactions, since a gauge boson mass violates gauge invariance, as we pointed out for the gauge quanta $\boldsymbol{b}_{\mu}$ in Section 6.2. One could try, then, to settle for a theory involving massive bosons without it being a gauge theory. Unfortunately, such a theory would not be renormalisable ([A1]). Let us explain briefly what this means.

When a theory is used to calculate the effects of fundamental forces at the quantum level, the obtained values are in certain cases infinite ([288]). Of course, if one uses a theory to calculate an observable quantity, and finds that the answer is infinite, one concludes either that a mathematical mistake has been made, or that the original theory was no good. For example, the electron's ability constantly to emit and reabsorb 'virtual' photons means that its total energy and its mass are infinite ([289]). However, by redefining the mass of the 'bare' electron to include these virtual processes and setting it equal to the measured mass – that is, by renormalising – the problem is removed. More generally, renormalisation is a procedure in quantum field theory by which divergent parts of a calculation, leading to nonsensical infinite results, are absorbed by redefinition into a few measurable quantities, so yielding finite answers. A relatively limited number of theories are renormalisable, with gauge theories representing a notable subset. ([290])

One strategy adopted to fix the mass problem of Yang-Mills bosons was to artificially endow the bosons with a mass greater than zero. Imposing a mass on the bosons, results in a finite range of the fields. If the mass is large enough, the range can be made as small as is wished. However, with this modification the local symmetry of the Yang-Mills theory would no longer be exact but approximate.

It turned out that the modified Yang-Mills theory had the problem of infinities and the standard renormalisation procedure used in QED did not work ([291]). An important idea was introduced in 1963 by Feynman in order to mitigate this problem. It is the notion of a *ghost particle*. Such a particle is added to a theory in the course of calculation and then vanishes when the calculation is finished. The use of a ghost particle can be justified if it never appears in the final state. This can be ensured by making certain the total probability of producing a ghost particle is always zero. ([H8])

This line of research was pursued by Martin J.G. Veltman (the University of Utrecht) and John S. Bell (CERN) ([292]). However, Veltman managed to renormalise his theory up to Feynman

---

[288] When quantising a classical field theory, such as electromagnetism, the trick is to break the field down into a sum of harmonic oscillators applying a Fourier-transform. These oscillators can then be quantised. However, when decomposing a field into elementary quanta, there are infinitely many possibilities, and each of these has a nonzero energy in its 'ground state' (lowest energy state). An infinite number of nonzero things added together can yield infinity ([T5]).

[289] These divergences seemed unavoidable consequences of locality (point-like particles with contact interactions) and unitarity (conservation of probabilities). Indeed, one must sum over the contribution of virtual photons with arbitrarily high energies because there is no short-distance structure. And due to conservation of probabilities, all processes contribute additively.

[290] Renormalisation was first developed in quantum electrodynamics (QED) to make sense of infinite integrals in perturbation theory. It was initially viewed as a suspect provisional procedure.
F. Dyson once asked Dirac: *"Well, Professor Dirac, what do you think of these new developments in quantum electrodynamics?"* Dirac, the mathematical purist, was not enamoured: *"I might have thought that the new ideas were correct if they had not been so ugly."* ([F1]).
Freeman John Dyson (1923 – 2020) was a British-American theoretical and mathematical physicist.

[291] QED is a well-behaved theory in which the infinities can be unambiguously removed by introducing two parameters that have to be determined experimentally: the mass and charge of electrons. ([H12]) Roughly speaking, a theory is renormalisable if it requires only a finite number of counter terms for canceling infinities.

[292] Martinus Justinus Godefriedus 'Tini' Veltman (1931 – 2021) was a Dutch theoretical physicist. He shared the 1999 Nobel Prize in Physics with his former PhD student Gerardus 't Hooft.
John Stewart Bell (1928 – 1990) was a physicist from Northern Ireland and the originator of Bell's theorem, an important theorem in quantum physics regarding hidden variable theories.

diagrams with one loop only. To render the 'massive Yang-Mills theory' renormalisable, a better theory was needed. ([H8])

In the meantime another new ingredient for the formulation of gauge theories was introduced. A way was found of giving gauge field quanta a mass, which is by 'spontaneously breaking' ([293]) the gauge symmetry. This means the symmetry exists at high energy and spontaneously breaks at lower energies. ([H8])

The solution is the so-called *Brout-Englert-Higgs mechanism* ([294]), or shortly BEH mechanism, that assigns a mass to the gauge bosons without breaking the gauge invariance. It requires the introduction of four new spin-0 fields, which are called *Higgs fields* (P. Higgs 1964, R. Brout and F. Englert 1964, G. Guralnik, C. R. Hagen and T. Kibble 1964 ([295])).

Spontaneous symmetry breaking (SSB) process can describe systems where the Lagrangian density obeys symmetries, but the lowest-energy vacuum solutions do not exhibit that same symmetry. When the system goes to one of those vacuum states, the symmetry is broken even though the entire Lagrangian density retains that symmetry. It means that the Lagrangian density can have symmetries that no longer appear in the physically realized state.

This process of spontaneous symmetry breaking was initially studied as a purely theoretical idea of endowing some of Yang-Mills fields with mass while retaining exact gauge symmetry. This idea did not gain much support or attention at that time. It is worth mentioning that the manuscript of the second paper of Higgs was firstly submitted to Physics Letters. After rejection by the editor, he resubmitted it to Physical Review Letters and got published due to the encouragement and support of Yoichiro Nambu ([296]) who was the referee of the paper.

One problem with the idea of spontaneous symmetry breaking was an earlier theorem of J. Goldstone [G5] ([297]) who proved in 1961 that whenever a (global) continuous symmetry is spontaneously broken by the vacuum state of a model, there must exist a massless spin-0 particle (called now *Goldstone boson*). Since there was no evidence for such particles, the spontaneous breaking of symmetries was considered to be unviable for a couple of years. ([298])

---

[293] The name is a bit misleading because the symmetry is not really broken. After all, the field equations and the Lagrangian are still invariant.

[294] This mechanism is simply called the *Higgs mechanism*. Actually it should be called Brout-Englert-Guralnik-Hagen-Higgs-Kibble-Mechanism. There is an anecdote about naming of concepts in physics [W13]: *"... In a course on particle physics I took at Harvard from (...) Alvaro De Rujula, whenever he introduced a concept with someone's name attached to it, he would generally say something like the following: 'This is the so-called Weinberg angle, which of course was discovered not by Weinberg, but by Glashow'. On one occasion after introducing a named concept he stopped for a while and seemed to be thinking deeply. Finally he announced that, as far as he knew, strangely enough, this concept actually seemed to have been discovered by the person whose name was attached to it".*

[295] - Robert Brout (1928 – 2011) was a Belgian theoretical physicist.
- François, Baron Englert (1932 –) is a Belgian theoretical physicist and 2013 Nobel Prize (shared with Peter Higgs) laureate.
- Peter Ware Higgs (1929 – 2024) is a British theoretical physicist and 2013 Nobel Prize laureate.
- Gerald Stanford 'Gerry' Guralnik (1936 – 2014) was an American theoretical physicist.
- Carl Richard Hagen (1937 –) is a professor of particle physics at the University of Rochester.
- Sir Thomas Walter Bannerman Kibble (1932 – 2016) was a British theoretical physicist.
While widely considered to have authored the most complete of the early papers on the Higgs theory, Guralnik+Hagen+Kibble were controversially not included in the 2013 Nobel Prize in Physics.

[296] Yoichiro Nambu (1921 – 2015) was a Japanese-American physicist and 2008 Nobel Prize (shared with Makoto Kobayashi and Toshihide Maskawa).

[297] Jeffrey Goldstone (1933 –) is a British theoretical physicist.

[298] In 1965 Higgs received an invitation from Freeman Dyson to present a seminar on the Higgs mechanism at the Institute for Advanced Study. When he delivered the seminar in March 1966, the audience was sceptical, with one Harvard theorist later admitting that they "*had been looking forward to tearing apart this idiot who thought he could get around the Goldstone theorem*." ([B1]).

In 1964 Higgs [H6] showed, however, that the theorem of J. Goldstone does not apply to gauge theories.

We begin by explaining the BEH mechanism in the simplest setting known as the Abelian Higgs model. This is the U(1) gauge theory (see Section 7.2) with the spontaneous symmetry breaking phenomenon to render the U(1) gauge boson massive. The key idea of the BEH mechanism is to include in the theory an extra field, one having the peculiar property that it does not vanish in the vacuum. In physics, the vacuum is defined as the state in which all fields have their lowest possible energy. For most fields the energy is minimized when the value of the field is zero everywhere, The Higgs field has this unusual property that reducing it to zero costs energy – its energy is smallest when the field has some value greater than zero. ([H8])

Brout et al. essentially studied spontaneously breaking the U(1) gauge symmetry by including one complex scalar field $\phi = \phi_1 + \mathrm{i}\phi_2$, acquiring a non-zero vacuum expectation value (VEV). Let us look at this more closely (see e.g. [A1], [M1], [N1], [S11], [T1] for further details). To understand the effect of SSB on a theory with a local symmetry we consider a toy model of 'electrodynamics' specified by the Lagrangian density

(7.33) $$\mathcal{L} = D_\mu\phi^\dagger D^\mu\phi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} - V(\phi),$$ ([299])

where ϕ is a complex scalar field, A_μ is a vector gauge field,

$$D_\mu = \nabla_\mu + \mathrm{i}gA_\mu$$ ([300]),

and the potential $V(\phi)$ is given by

$$V(\phi) := \mu^2\phi^\dagger\phi + \lambda(\phi^\dagger\phi)^2,$$

where $\mu^2, \lambda > 0$. Notice that V is also a function of $|\phi|^2$, because $\phi^\dagger\phi = |\phi|^2$. The Lagrangian density (7.33) is invariant under U(1) gauge transformation (5.14). Since μ^2 and λ are positive, the potential $V(\phi)$ has a minimum at $\phi = 0$. We call the $\phi = 0$ state the vacuum. In terms of a quantum field theory, where $\hat{\phi}$ is an operator, the precise statement is that the operator $\hat{\phi}$ has zero vacuum expectation value (VEV), i.e. $\langle\hat{\phi}\rangle = \langle 0|\hat{\phi}|0\rangle = 0$ (see Section 5.9 for notation).

Now suppose we reverse the sign of μ^2, so that the potential becomes

$$V(\phi) = -\mu^2\phi^\dagger\phi + \lambda(\phi^\dagger\phi)^2.$$

We see that this potential no longer has a minimum at $\phi = 0$, but a (local) maximum. The minimum occurs at

(7.34) $$|\phi|^2 = \frac{\mu^2}{2\lambda} = \frac{v}{\sqrt{2}},$$

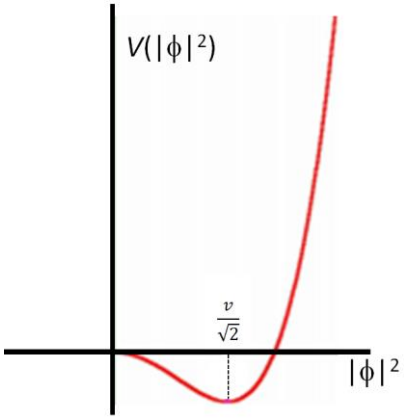
where $v =: \frac{1}{\sqrt{2}}\frac{\mu^2}{\lambda}$. The potential V as a function of $|\phi|^2$ ([301]):

---

[299] Reminder: $F_{\mu\nu} = \nabla_\mu A_\nu - \nabla_\nu A_\mu$ – see Example 7.7.

[300] We call the gauge coupling constant g rather than q (like in (5.8)) because we are using this theory as a formal example rather than a physical model.

[301] In 3D the Higgs potential V(ϕ) has the rotationally symmetric shape of a 'Mexican hat' – see below.



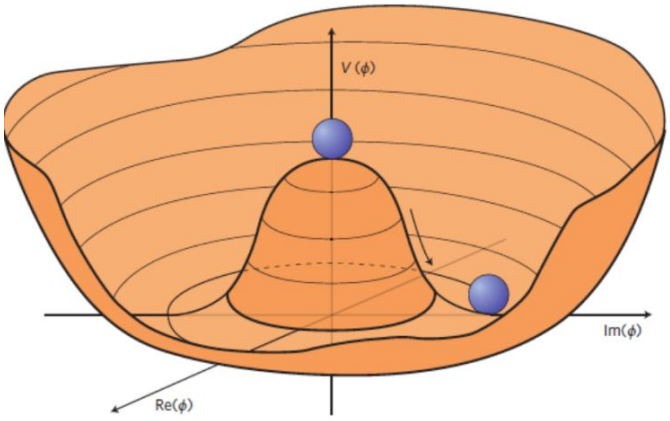
Thus the potential V(ϕ) attains its minimum value at

$$\phi = e^{i\theta}\phi_0,$$

where θ can take any value from 0 to 2π and ϕ_0 is such that $|\phi_0|^2 = \frac{v}{\sqrt{2}}$.

Consequently, there is an infinite number of states each with the same lowest energy, i.e. we have a degenerate vacuum. And the system is completely symmetric: one can rotate the potential around the vertical axis (by taking different values of θ), and it will still look exactly the same. The symmetry breaking occurs in the choice made for the value of θ which represents the true vacuum, i.e. the choice of a particular point the field 'rolls down' into. For convenience, one chooses $\theta = 0$ to be vacuum. Such a choice constitutes a spontaneous breaking of the U(1) invariance since a U(1) transformation (5.14) takes us to a different point with the lowest energy state. In other words, the vacuum breaks U(1) invariance.

When rolling down into a point of lower energy, the field ϕ acquires a nonzero vacuum expectation value (VEV). Previously, when it was on the top of the hill, the field's value was zero. Now the field has a non-zero value, the VEV $= \frac{v}{\sqrt{2}}$, but it has lower energy than it had before. ([302]) Note that the VEV is the value of the field, not of the energy. The potential V as a function of ϕ:



([303])

In order to understand the physical content of the theory we can write the complex field ϕ in terms of its modulus $|\phi|$ and a phase, and to expand $|\phi|$ around VEV $= v/\sqrt{2}$,

---

[302] It is a very unique property of the field ϕ, however. Most fields are in their lowest state of energy when they are free of excitations, that is, no particles are present. For instance, the lowest energy state of the electromagnetic field is when there is no radiation, i.e., there are no photons present. It is not the case for the ϕ field. Its lowest energy state is when some excitations are present. So the 'no excitations present' state is unstable: it can 'decay' into a lower energy state. This is the process of symmetry breaking which creates a new type of vacuum with the vacuum expectation value (VEV) being nonzero.

[303] Spontaneous symmetry breaking: an object residing in a rotationally symmetric potential rolls down and finds a stable, asymmetric position. From https://cds.cern.ch/record/2012465/plots

$$\phi = |\phi| e^{i\chi/v} = \frac{1}{\sqrt{2}}(v + H) e^{i\chi/v},$$

where χ is a real scalar field. Thus we have separated the field ϕ into VEV, which is just a number, and a new real scalar field $(t, \boldsymbol{x}) \to H(t, \boldsymbol{x})$. Substituting it into the Lagrangian (7.33), we find

$$\begin{aligned}\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} - g v A_\mu \nabla^\mu \chi + \frac{1}{2}(gv)^2 A_\mu A^\mu \\ + \frac{1}{2}\left(\nabla_\mu H \nabla^\mu H + 2\mu^2 H^2\right) \\ + \frac{1}{2}\nabla_\mu \chi \nabla^\mu \chi + (H, \chi \text{ interactions}).\end{aligned}$$

This Lagrangian now describes a theory with a 'photon' of mass $m_A = gv$, a Higgs (i.e. Brout-Englert-Higgs) boson H with $m_H = \sqrt{2}\mu = \sqrt{2\lambda}v$, and a massless Goldstone boson χ. However, the Goldstone boson can be removed by making the following gauge transformation:

$$A_\mu \to A'_\mu = A_\mu - \frac{1}{gv}\nabla_\mu \chi.$$

The gauge choice with the transformation above is called the *unitary gauge*. The Goldstone boson will then completely disappear from the theory and one says that the Goldstone has been 'eaten' to give the photon mass.

7.4. The Glashow-Salam-Weinberg electroweak theory (GSW)[304]

The decay of nuclei proceeds via α-, β-, γ-emission, or by fission. The first is governed by strong nuclear forces, the third by electromagnetic interactions. The β-emission has an entirely different origin. It is caused by *weak interactions*.

The weak interaction is in a certain sense the most enigmatic. The two macroscopic interactions, gravitation and electromagnetism, are both familiar to us, at least in the classical version. The strong nuclear force is absolutely essential for the existence of matter as we know it. But what do we 'need' the weak interaction for? It is of extremely short range and at low energies by far the weakest of the three microscopic interactions. But actually also the weak interaction is vital for life on our planet. Without it the production of energy in the sun would not be possible because the weak interaction plays an essential role in nuclear fusion responsible for energy production. ([E1])

Initially, β-emission posed a serious puzzle: the observed emerging particles, electrons or positrons, did not come out with definite energy equal to the mass difference between initial and final nuclei. Instead, they had an energy distribution with a maximal energy at the expected value. This led Niels Bohr ([305]) to speculate that energy conservation could possibly be violated in quantum physics. This shows how desperate people were. The solution to the problem was devised by Wolfgang Pauli. In 1930 he suggested in a historic letter [P4] to colleagues ("Dear Radioactive Ladies and Gentlemen") attending a meeting at Tübingen that the missing energy was carried off by a neutral particle, now called *neutrino* ([306]) (excerpt):

---

[304] See [A1] and [H9] as a general reference for this and the next section.

[305] Niels Henrik David Bohr (1885 – 1962) was a Danish physicist who made foundational contributions to understanding atomic structure and quantum theory, for which he received the Nobel Prize in Physics in 1922.

[306] Neutrinos feel only the weak interaction, which is what makes them so difficult to study. They are the only particles to experience just one of the fundamental forces. Pauli writes in his letter that he does not dare to publish his idea for the time being. Half a year later, at a meeting of the American Physical Society in Pasadena, California in July 1931, Pauli himself presented his neutrino hypothesis but he still prohibited any publication. Only his talk at the 7th Solvay Conference in Brussels in 1933 could finally be published. ([E1])

Offener Brief an die Gruppe der Radioaktiven bei der Gauvereins-Tagung zu Tübingen.

Abschrift

Physikalisches Institut
der Eidg. Technischen Hochschule
Zürich

Zürich, 4. Dez. 1930
Gloriastrasse

Liebe Radioaktive Damen und Herren,

Wie der Ueberbringer dieser Zeilen, den ich huldvollst anzuhören bitte, Ihnen des näheren auseinandersetzen wird, bin ich angesichts der "falschen" Statistik der N- und Li-6 Kerne, sowie des kontinuierlichen beta-Spektrums auf einen verzweifelten Ausweg verfallen um den "Wechselsatz" (1) der Statistik und den Energiesatz zu retten. Nämlich die Möglichkeit, es könnten elektrisch neutrale Teilchen, die ich Neutronen nennen will, in den Kernen existieren,

(...)

Weak interactions act on all particles (except photons and gluons), both leptons and quarks. Furthermore, the weak interaction is the only one that can change the flavour of the particles involved in the interaction (see Remark 2.2 in Chapter 2). Despite the ubiquity and importance of the weak interaction, though, weak processes are relatively rare. We observe them taking place in a system only if strong or electromagnetic processes are forbidden for some reason, e.g. by conservation laws.

We know now there are three mediating spin-1 bosons for the weak interaction: the W^+ and W^- carry an electric charge of 1 elementary charge and are each other's antiparticles. The Z^0 boson is electrically neutral and is its own antiparticle. These bosons are among the heavyweights of the fundamental particles. With masses of 80.4 GeV and 91.2 GeV, respectively, the $W^\pm$ and Z^0 bosons are almost 80 times as massive as the proton – heavier, even, than entire iron atoms. This is the reason why the weak interaction seems so feeble. The uncertainty principle tells us that in order for two particles to be close enough to exchange a virtual particle with a mass of the order of 100 GeV, and thus participate in a weak interaction event, they have to be very close together – about 10^{-17} metres apart, or roughly a hundredth the radius of a proton. The intrinsic strength of the weak interaction is not small at all. In fact, the weak coupling constant is $1/29$ – almost five times larger than the (electromagnetic) fine-structure constant $\alpha \approx 1/137$ ([307]).

The emission or absorption of a $W^\pm$ boson can change the electric charge and flavour (but not colour) of the particle – for example changing a strange quark into an up quark. The emission or absorption of a Z^0 boson can only change the spin, momentum, and energy of the other particle.

Looking for a model of the weak interactions based on a version of the Yang-Mills theory the question is this: what is the relevant symmetry group of local phase transformations, i.e. the relevant weak gauge group? Several possibilities were suggested, but it is now very well established that the one originally proposed by Glashow (1961), subsequently treated as a spontaneously broken gauge symmetry by Weinberg (1967) and by Salam (1968) ([308]), and

[307] This name comes from the fact that it was first measured in the splitting of atomic spectral lines called 'fine structure'.

[308] Sheldon Lee Glashow (1932 –) is an American theoretical physicist and 1979 Nobel Prize (shared with Weinberg and Salam) laureate.
Steven Weinberg (1933 – 2021) is an American theoretical physicist and 1979 Nobel Prize laureate.
Mohammad Abdus Salam (1926 – 1996) was a Pakistani theoretical physicist and 1979 Nobel Prize laureate.

later extended by other authors, produces a theory which is in remarkable agreement with currently known data.

The simplest possibility is to suppose that the relevant symmetry group is 'weak SU(2) group' called *weak isospin*. Let us emphasize that this weak isospin is completely different from the isospin of Section 6.1. The only agreement between these two notions is that they are described by the same mathematics of the SU(2) symmetry group. But now, our doublet is not (n, p) but rather lepton doublets, e.g. (ν_e, e^-) – the electron e^- and its neutrino ν_e, or quark doublets, e.g. (u, d). The other lepton doublets are (ν_μ, μ) and (ν_τ, τ), whereas quark doublets are (c, s) and (t, b) – see Chapter 2.

Taking the SU(2) symmetry group looked like a good model for weak interactions, but physicists encountered its first big flaw. It was observed in the 1950s by C.S. Wu and collaborators (at the suggestion of Lee and Yang) that the weak interactions did not conserve parity ([309]). That is to say, the Lagrangian is not invariant under the spatial inversion $\boldsymbol{x} \to -\boldsymbol{x}$. Particle physicists observed only left-chiral fermions and right-chiral antifermions ([310]) engaging in the charged weak interaction. Interactions involving right-chiral fermions have not been shown to occur, implying that the universe has a preference for left-chirality. This is a striking observation, since parity is a symmetry that holds for all other fundamental interactions ([W11]).

We shall examine the concept of parity in more detail. The parity operation is defined as spatial inversion around the origin:

$$(t, \boldsymbol{x}) \to (t, -\boldsymbol{x}) = (t, -x, -y, -z).$$

In other equivalent terms, the parity operation is carried out by taking the mirror image of an object and then rotating it by 180° around the axis perpendicular to the mirror. This operation represents a symmetry in space. It may be anticipated that the laws of nature would remain unaltered under this operation, given that a preferred direction in space does not exist.

The *parity operator* $\hat{P}$ transforms a wave function $(t, \boldsymbol{x}) \to \psi(t, \boldsymbol{x})$ to

$$\psi'(t, \boldsymbol{x}) = \hat{P}\psi(t, \boldsymbol{x}) := \psi(t, -\boldsymbol{x}).$$

The operator $\hat{P}$ is both unitary and Hermitian, and thus corresponds to an observable quantity. If ψ is an eigenfunction of the parity operator with eigenvalue P then $P = \pm 1$. The parity of a fermion is opposite that of the antifermion, whereas the parity of a boson is the same as its antiboson. It is customary to assign *positive* or *even* parity $+1$ to particles and *negative* or *odd* parity -1 to antifermions (if they are fermions). The parity of a combined system is the product of the parity of its constituent parts. ([B6], [M7])

An explicit form for the parity operator, suitable for use on Dirac spinors, is ([L1])

$$\hat{P} := \gamma_0 = \begin{bmatrix} 0_2 & I_2 \\ I_2 & 0_2 \end{bmatrix}.$$

It turns a left-chiral spinor into a right-chiral spinor and vice versa: $\hat{P}\psi_L \to \psi_R$ and $\hat{P}\psi_R \to \psi_L$.

---

[309] Since 1925, physicists had assumed that our world is indistinguishable from its mirror image – a notion known as *parity conservation*. In June of 1956 theoretical physicists Tsung Dao Lee and C.N. Yang studied the so-called θ-τ meson puzzle and submitted a short paper to the Physical Review raising the question of whether parity is conserved in weak interactions. They proposed a number of experimental tests for parity conservation in the weak interaction, and in the same year such an experiment was carried out by C.S. Wu which confirmed parity violation in the weak interaction. Yang and Lee were awarded the Nobel Prize in Physics in 1957. The success of the experiment came as a great surprise to Pauli and Feynman. They had been willing to bet money that the experiment would find that parity is preserved. Pauli: "*After all, God is not a weak left-hander*" ([E1]). Feynman considered parity violation "*unlikely, but possible, and a very exciting possibility*," but later made a 50 $ bet with a friend that parity would not be violated. Although Feynman had lost the bet, he was among the first to draw the right conclusions. Feynman and Gell-Mann modified the original Fermi theory of weak interactions to account for parity violation (V-A theory, 1958). ([E1])

[310] See Section 5.7 for explanation of 'chirality'.

The term 'parity violation' is used to describe the phenomenon whereby the weak force treats a quantum and its parity reverse (its other chirality in the case of fermions) differently. This violation of parity has been observed on numerous occasions. The first measurement that demonstrated such a violation of parity was conducted in 1956 by Chien-Shiung Wu in collaboration with the Low Temperature Group of the US National Bureau of Standards.

C.S. Wu devised and conducted an experiment to test the possibility of parity violation in β-decay. She set up a system of Cobalt-60 atoms, which all decayed to Nickel-60. She aligned them in a magnetic field, so that all their spin vectors were lined up, and then let them to decay, measuring the direction of the outgoing electron. If parity were conserved, she would expect to see electrons emitted isotropically. For what reason? The parity operator has no effect on the spin state of a cobalt atom. This implies that the spin state is the same in both the original world and the parity-transformed mirror world. Let us consider the scenario in which an electron is emitted in the direction of the spin vector in this world. In the mirror world, the electron will be moving in the opposite direction to that of the spin. The principle of parity conservation implies that the probability of one interaction occurring in this world is the same as the probability of its mirror image. Consequently, we should observe the same number of events in which the electron is emitted anti-parallel to the spin vector as in which it is emitted parallel to the spin vector. ([B6])

However, Wu observed that electrons were emitted with a preferential directionality aligned with the spin vector, a phenomenon that clearly violated the conservation of parity. Furthermore, the emission was not insignificant, with the vast majority of electrons being emitted in a single direction. It appeared that the violation was at its most extreme. The long-held belief that parity was a fundamental symmetry of nature was challenged in 1956, leading to significant distress among many respected physicists. ([B6]).

The question thus arises as to how parity violation can be incorporated into the model of weak interactions. The requirement of Lorentz invariance imposes significant constraints on the form of the interaction vertex ([311]). Both quantum electrodynamics (QED) and quantum chromodynamics (QCD) conserve parity and are vector interactions. Consequently, the vertex can be expressed as $j_\mu = \overline{\psi}\gamma_\mu\phi$. Given that parity is violated in the weak interaction, it can be concluded that the weak interaction vertex cannot be expressed in this form. The form of the weak interaction is determined by experiment to be vector – axial-vector (V–A) interaction ([312]). The charged weak current vertex involves a chirality projection and is written as

$$j_W^{CC} = \frac{g}{\sqrt{2}}\overline{\psi}\gamma_\mu\frac{1}{2}(1-\gamma_5)\phi,$$

where g is the weak coupling constant. The name 'charged current' comes from the currents of fermions coupled to the $W^\pm$ bosons, which have an electric charge. ([B6])

The incorporation of the left-chiral projection operator in the current implies that the charged weak interaction only couples left-chiral fermions, or right-chiral antifermions ([B6])

$$\overline{\psi}\gamma_\mu\frac{1}{2}(1-\gamma_5)\phi = (\overline{\psi_L}+\overline{\psi_R})\gamma_\mu\phi_L = \overline{\psi_L}\gamma_\mu\phi_L.$$

At extremely high energies, the chiral components correspond to helicity eigenstates (see Section 5.8) ([313]). This has implications for neutrinos. It is established that all neutrinos are observed to possess left-handed (i.e. negative) helicity, whereas anti-neutrinos exhibit right-handed (i.e. positive) helicity. Given that neutrinos, even if they do possess mass, are ultra-relativistic, this implies that all neutrinos have left-chirality, whereas antineutrinos have right-

[311] A point in a Feynman diagram where lines connect to other lines is a *vertex*, and this is where the particles meet and interact: by emitting or absorbing new particles, deflecting one another, or changing type ([W11]).

[312] An axial vector (also known as a *pseudovector*) is a vector quantity that remains unchanged when transformed according to a parity transformation.

[313] Recall that the concept of chirality does not apply to bosons. With regard to helicity, it has been established that in the rest frame of the collision, the $W^\pm$ and Z^0 bosons are produced in all three helicity states, with left-handed helicity being the predominant outcome ([C4], [Q1]).

chirality. As neutrinos can only be produced through weak interactions, they are all created as left-chiral particles. ([B6])

It is important to note that neutrinos do not inherently possess intrinsic left-handed helicity. The left-handed helicity of neutrinos is a consequence of their creation in weak interactions and the fact that, due to the negligible mass of neutrinos, helicity and chirality are essentially synonymous in this case. This does not preclude the possibility of the existence of a neutrino with right-handed helicity. Nevertheless, it can be demonstrated that the probability of generating such a neutrino is extremely low. However, if a right-handed helicity neutrino exists, it does not couple to any of our fundamental forces (with the possible exception of gravity) and therefore may be very difficult to detect. Despite this, due to the phenomenon of neutrino oscillations, a right-handed helicity neutrino state could still have indirect but visible effects in some neutrino oscillation experiments. ([B6])

In the case of electrons (and quarks), however, both left- and right-chiral states have been observed, yet only the former couple to the charged weak interaction, i.e. to the $W^{\pm}$ bosons ([314]). In contrast, the Z^0 boson is capable of coupling also to right-chiral particles. Given that neutrinos are created solely by the charged weak current, this has no impact on the properties of the neutrino. ([B6])

In conclusion, the weak force acts on all quarks and leptons. It is propagated by three massive bosons: W^+, W^- and Z^0. Interactions of charged bosons (known as 'charged weak current interactions') change the flavour of a (left-chiral) fermion, whereas Z^0-boson interactions (termed 'neutral weak current interactions') maintain the flavour of the fermion.

Let us now resume our discussion of the GSW theory of weak interactions.

Since the weak charged bosons couple only to left-chiral fermions (or their right-chiral antiparticles), the weak isospin group is referred to as $\mathrm{SU}(2)_L$, to show that the weak isospin assignments and corresponding transformation properties apply only to these left-chiral parts.

Weak isospin is a quantum number related to the weak interaction. It is usually given the symbol T with the third component written as T_3 (this is analogous to the spin component S_z in Example 5.3). Left-chiral fermions and right-chiral antifermions have a total spin of $T = {}^1\!/_2$ and can be grouped into doublets with $T_3 = \pm {}^1\!/_2$, which exhibit identical behaviour under the weak interaction. By convention, electrically charged fermions are assigned T_3 with the same sign as their electric charge. Right-chiral fermions and left-chiral antifermions possess zero weak isospin $T = 0$, rendering them incapable of interacting with the $W^{\pm}$ bosons (with the exception of electrical interaction). Consequently, for example, the right-chiral electron undergoes a trivial transformation (it remains unchanged) under SU(2), while the left-chiral electron transforms into the neutrino and vice versa.

For a fermion field ψ let ψ_{L} denote its left-chiral component (it is a Weyl spinor – see Section 5.8). An $\mathrm{SU}(2)_L$ transformation (6.8) acts now on left–chiral parts of doublets, e.g.

$$\begin{bmatrix} \nu_e \\ e^- \end{bmatrix}_L \rightarrow \begin{bmatrix} \nu_e \\ e^- \end{bmatrix}_L' = e^{\mathrm{i}\frac{1}{2}\boldsymbol{\alpha}\cdot\boldsymbol{\tau}} \begin{bmatrix} \nu_e \\ e^- \end{bmatrix}_L. \tag{7.35}$$

Making (7.35) into a local phase invariance (following the logic of Section 6.2) will entail the introduction of three gauge vector fields $\boldsymbol{W}_{\mu} = (W_{\mu}^1, W_{\mu}^2, W_{\mu}^3)$, transforming under the group $\mathrm{SU}(2)_L$. The covariant derivative is then given by (see Example 7.8)

$$D_{\mu} = \mathrm{I}_2\nabla_{\mu} - \mathrm{ig}\tfrac{1}{2}\boldsymbol{\tau}\cdot\boldsymbol{W}_{\mu} = \mathrm{I}_2\nabla_{\mu} - \mathrm{ig}\tfrac{1}{2}\textstyle\sum_{j=1}^{3}\tau_j W_{\mu}^j =$$

$$= \begin{bmatrix} \nabla_{\mu} & 0 \\ 0 & \nabla_{\mu} \end{bmatrix} - \mathrm{ig}\tfrac{1}{2}\begin{bmatrix} W_{\mu}^3 & W_{\mu}^1 - iW_{\mu}^2 \\ W_{\mu}^1 + iW_{\mu}^2 & -W_{\mu}^3 \end{bmatrix}$$

[314] A massive fermion, which oscillates between left- and right-chiral states, may only emit a $W^{\pm}$ particle (for instance, an electron turning into a neutrino) when it is in its left-chiral state.

$$= \begin{bmatrix} \nabla_\mu & 0 \\ 0 & \nabla_\mu \end{bmatrix} - \mathrm{i}g\frac{1}{2}\begin{bmatrix} W_\mu^3 & \sqrt{2}W_\mu^+ \\ \sqrt{2}W_\mu^- & -W_\mu^3 \end{bmatrix},$$

where we have defined a complex gauge field $W_\mu^\pm := \frac{1}{\sqrt{2}}(W_\mu^1 \mp iW_\mu^2)$ ([315]). The $\pm$ superscript on $W_\mu^\pm$ is just the electric charge (which is conserved) carried by the gauge boson. The charged quanta of these fields will, of course, be related to the now familiar physical $W^\pm$ bosons mediating the charged current transitions. The neutral member W_μ^3 corresponds to Z^0 that mediates neutral current weak interactions ([M6]). ([316])

However, the attempt to treat the weak interactions as a gauge theory with the symmetry group SU(2) initially failed. By (7.35) local SU(2) transformations act on doublets that are two component objects. Such a doublet ψ contains two spin-½ fields, for example, the electron and the electron neutrino field that are 'rotated' by SU(2) transformations into each other. A locally SU(2) invariant total Lagrangian density is – see Eq. (6.19) and Example 7.8

$$\mathcal{L}_{SU(2)} = \overline{\psi}(\gamma_\mu\nabla^\mu - m)\psi - \frac{1}{4}\boldsymbol{F}_{\mu\nu}\boldsymbol{F}^{\mu\nu} - \mathrm{i}g\,\overline{\psi}\gamma_\mu\boldsymbol{\tau}\cdot\boldsymbol{W}^\mu\psi,$$

We know that local SU(2) symmetry can only be achieved without 'mass terms' of the form $m\boldsymbol{W}_\mu\boldsymbol{W}^\mu$. Moreover, the Lagrangian density $\mathcal{L}_{SU(2)}$ would have to be invariant with some arbitrary 2×2 mass matrix M (see Example 7.8), because ψ is now a two-component object. Taking only one mass value m means that masses for the two spin-½ fields must be equal. But from experiments we know that this is not the case: the electron mass is much bigger than the electron-neutrino mass. This is commonly interpreted as the SU(2) symmetry being broken. As we shall see below, the BEH mechanism enables us to get a locally SU(2) invariant Lagrangian density that includes mass terms.

It turned out that the electromagnetic and the weak interactions cannot be treated by separate gauge theories. A key contribution was made by Glashow (1961); similar ideas were also advanced by Salam and Ward (1964). Glashow suggested enlarging the $SU(2)_L$ schemes by inclusion of an additional $U(1)_Y$ gauge group, resulting in an $SU(2)_L \times U(1)_Y$ group structure ([317]). The Abelian $U(1)_Y$ group is associated with a weak analogue of hypercharge – weak hypercharge Y – just as $SU(2)_L$ is associated with weak isospin (see Example 7.7). The mathematics of electric charge q and weak hypercharge Y is the same, but the physical meaning is slightly different.

Electroweak symmetry breaking in the early universe had an impact on the U(1) symmetry. So before symmetry breaking, the charge was different. This charge is *weak hypercharge* Y. Another consequence is that the gauge field was 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* ([S0]).

Naively one could think of U(1) as the group $U(1)_{em}$ for electromagnetism and at a SU(2) for weak (charged and neutral) interactions, but this does not work, because weak charged currents couple only to left-chiral components, while electromagnetic interaction acts on both left- and right-chiral components with the same strength. This means that the gauge boson of the U(1) group cannot be the photon. For this reason the B boson of a gauge field B_μ was introduced that couples to fermions according to the weak hypercharge Y. The field B_μ is analogous to the photon field A_μ in the $U(1)_{em}$ gauge theory (see Example 7.7). The weak hypercharge of the

---

[315] $\sqrt{2}$ is included for convenience.

[316] The physical boson Z^0 is related to a mixture of W_μ^3 and B_μ gauge fields – see below.

[317] In group theory, the (direct) product is an operation that takes two groups G and H and constructs a new group, usually denoted G × H. That is, G × H is the set of all ordered pairs (g, h), where g ∈ G and h ∈ H, and the group multiplication is defined component-wise: $(g_1, h_1)\cdot(g_2, h_2) = (g_1g_2, h_1h_2)$.

left-chiral electron and the neutrino is $Y = -1$ ([318]). The right-chiral electron has a weak hypercharge $Y = -2$. However, we cannot measure the weak hypercharge since we cannot create the circumstances from before symmetry breaking in our detectors. So we have to rely on theory ([S0]).

Consequently, in the terminology of Section 7.2, Glashow studied chiral G-gauge theory with the group $G = SU(2)_L \times U(1)_Y$. Because there are two separate groups there are two different coupling constants g and g′ (see (7.39) below) corresponding to the two different interactions – so there is no complete unification in terms of coupling strengths. The first underlying problem with Glashow's idea is, of course, that if this G symmetry holds exactly, it demands that all the bosons have zero mass. This does not reflect the world as we see it. The second problem was more subtle. No weak interaction had been observed that would require the exchange of a new neutral particle Z^0. For these reasons, Glashow's proposal drifted into the background. Yet several of the key theoretical ingredients needed to complete a revolution in fundamental physics were in place, but it was far from obvious at the time.

It was Weinberg (1967) and Salam (1968) who made the correct application of the BEH mechanism of spontaneous symmetry breaking of $SU(2)_L \times U(1)_Y$ in order to generate mass for the gauge quanta associated with the weak force.

Let us briefly sketch that idea. Analogously to the U(1) toy model in Section 7.3, the electroweak $SU(2)_L \times U(1)_Y$ symmetry is spontaneously broken by introducing an additional field ϕ with an appropriate potential. In order to obtain three massive and one massless electroweak gauge bosons, the latter one being the photon, $SU(2)_L \times U(1)_Y$ must be broken to the electromagnetic group $U(1)_{em}$. Therefore, a Higgs mechanism must operate in such a way that after symmetry breaking one massless gauge boson (the photon) remains, and three others acquire a mass.

The local $SU(2)_L \times U(1)_Y$ symmetry requires three $SU(2)_L$ gauge fields (Example 7.8), which we call W_μ^i (i = 1, 2, 3), and one $U(1)_Y$ gauge field B_μ (Example 7.7). Now, the easiest way to obtain at least three scalar degrees of freedom is to introduce a complex scalar SU(2) field having four (real) degrees of freedom

$$\phi = \begin{bmatrix} \phi^+ \\ \phi^0 \end{bmatrix} = \begin{bmatrix} \chi_1 + i\chi_2 \\ H + i\chi_4 \end{bmatrix}. \tag{7.36}$$

The field ϕ has two neutral and two electrically charged components that form a complex doublet of the weak isospin SU(2) symmetry ([319]). One degree of freedom is Higgs mode H acquiring a non-zero VEV, the remaining three are massless Goldstone bosons χ_i with zero VEV. Choosing which one of the four degrees of freedom in the doublet (7.36) to be the Higgs mode is arbitrary. In addition, we need a scalar potential with the property that its ground state no longer preserves the symmetries of the theory. This will generate mass terms for both the weak gauge bosons and the (fundamental charged) fermions starting from a Lagrangian that preserves the symmetries of the theory.

Thus one must now decide how to choose the non-zero vacuum expectation value that breaks the $SU(2)_L \times U(1)_Y$ symmetry. The essential point is that, after symmetry breaking, we should be left with three massive gauge bosons (which will be the $W^\pm$ and Z^0) and one massless gauge boson, the photon γ ([320]). We may reasonably guess that the massless boson will be

---

[318] The weak hypercharge Y is related to the electric charge Q (in elementary charge units) and the third component of weak isospin T_3 by the Gell-Mann-Nishijima formula $Q = T_3 + \frac{1}{2}Y$. Thus for the left-chiral e^- (and neutrino) we have $Y = 2(Q - T_3) = 2(-1 - (-\frac{1}{2})) = -1$.

[319] This complex doublet model is just the simplest one. There are alternatives in which more than one Higgs boson exists, which means that there would be heavier Higgs-bosons that are yet to be found. ([T5])

[320] The rules of quantum mechanics allow the W_μ^i and B_μ bosons to mix to form $W^\pm$, Z^0 and γ. The amount of mixing is determined by a number θ_W called the *weak mixing angle* – see below.

associated with a symmetry that is unbroken by the vacuum expectation value. However, we want to generate mass for $W^{\pm}$ and Z^0. The choice suggested by Weinberg (1967) was

$$\phi_0 = \langle 0|\hat{\phi}|0\rangle = \begin{bmatrix} 0 \\ \frac{v}{\sqrt{2}} \end{bmatrix},$$

where v is defined as in Eq. (7.34). The parameter $v = 246$ GeV defines the electroweak scale, also known as the *Fermi scale*. The fluctuation around the minimum v is can be written as

$$\phi(\mathrm{t}, \boldsymbol{x}) = \phi_0 + \frac{1}{\sqrt{2}} H(\mathrm{t}, \boldsymbol{x}).$$

Thus

$$\phi = \begin{bmatrix} 0 \\ \frac{1}{\sqrt{2}}(v + H) \end{bmatrix} \tag{7.37}$$

Where $(\mathrm{t}, \boldsymbol{x}) \rightarrow H(\mathrm{t}, \boldsymbol{x})$ is the physical (real scalar) Higgs field ([321]). Its SU(2)-invariant Lagrangian density, consisting of a kinetic and a potential term, is

$$\mathcal{L}_{Higgs} = D_\mu \phi^\dagger D^\mu \phi + \mu^2 \phi^\dagger \phi - \lambda(\phi^\dagger \phi)^2 \quad (\mu^2, \lambda > 0). \tag{7.38}$$

The Lagrangian density for the sector containing the gauge fields and the Higgs fields is

$$\mathcal{L} = \mathcal{L}_{Higgs} + \mathcal{L}_{SU(2)} + \mathcal{L}_{U(1)_Y} = \mathcal{L}_{Higgs} - \frac{1}{4}\boldsymbol{F}_{\mu\nu}\boldsymbol{F}^{\mu\nu} - \frac{1}{4} G_{\mu\nu} G^{\mu\nu},$$

where

$$\boldsymbol{F}_{\mu\nu} := \nabla_\mu \boldsymbol{W}_\nu - \nabla_\nu \boldsymbol{W}_\mu + \mathrm{g}\boldsymbol{W}_\mu \times \boldsymbol{W}_\nu$$

is the SU(2) field strength tensor and $G_{\mu\nu}$ is the $\mathrm{U}(1)_Y$ field strength tensor – see Examples 7.6 and 7.5, respectively. The covariant derivative D_μ is given by

$$D_\mu = \nabla_\mu - \mathrm{ig}\boldsymbol{\tau} \cdot \boldsymbol{W}_\mu/2 - \mathrm{ig}' B_\mu/2, \tag{7.39}$$

where g′ is the coupling constant of the B_μ field (see Example 7.7). The Lagrangian density for the electroweak interactions before electroweak symmetry breaking is

$$\mathcal{L}_{GSW} = \mathcal{L}_{Higgs} + \mathcal{L}_{SU(2)} + \mathcal{L}_{U(1)_Y} + \mathcal{L}_f + \mathcal{L}_{Yukawa}, \tag{7.40}$$

where $\mathcal{L}_f$ is the kinetic term for the Standard Model fermions and $\mathcal{L}_{Yukawa}$ describes the Yukawa interaction with the fermions.

When the expansion (7.37) is inserted into the Lagrangian density (7.40), the original $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ symmetry is not apparent anymore and is said to be spontaneously broken ([322]). However, a $\mathrm{U}(1)_{em}$ symmetry remains, ensured by a vacuum expectation value in the neutral component of the scalar doublet (7.36) and not in the charged one. Therefore, the photon as unbroken U(1) generator remains massless. In contrast, the weak gauge bosons acquire masses from the kinetic term of (7.38). The vacuum expectation value generates quadratic terms for the $W^{\pm}$ and the Z^0 which are interpreted as mass terms $m_W^2 W_\mu^+ (W^-)^\mu$ and $m_Z^2 Z_{0\mu} Z_0^\mu$, i.e. those gauge bosons acquire masses: $m_W = m_Z \cos\theta_W = \frac{1}{2}\mathrm{g}v$, where θ_W is the weak mixing angle ([323]).

---

[321] The Higgs field is not charged under electromagnetism (it can not be, since it is real).

[322] One often reads that the term 'spontaneous symmetry breaking' SSB is misleading; the right term should be 'hidden symmetries', which refers to systems in which some symmetries of the law are not visible, i.e. hidden from the lowest energy solutions of the law equations. This seems to suggest that no symmetry is broken in such systems - rather different symmetries apply to different aspects of them. For a detailed discussion of this issue the reader is referred to [I3].

[323] The angle can be expressed in terms of the $\mathrm{SU}(2)_L$ and $\mathrm{U}(1)_Y$ coupling constants: $\cos\theta_W = \mathrm{g}/\sqrt{\mathrm{g}^2 + \mathrm{g}'^2}$. Because the value of the mixing angle is currently determined empirically, in the absence of any superseding theoretical derivation it is mathematically defined as $\cos\theta_W = m_W/m_Z$ ([W11]). In contrast to the charged current weak interaction $W^{\pm}$, the Z^0 boson couples to left-handed (LH) and right-handed (RH) chiral components, albeit with differing strengths.

Thus after symmetry breaking, charged leptons and quarks that interacted with the Higgs doublet field now interact with this new vacuum itself, through the nonzero Higgs VEV ≈ 246 GeV. ([324]) This shows up as an effective mass for these particles, because interactions mean energy which plays the role of rest mass in the equations.

One says that three of the four real degrees of freedom of the Higgs complex doublet (7.36) – the Goldstone bosons – are 'eaten' by the three vector bosons $W^{\pm}$ and Z^0, which acquire mass this way. Only one degree of freedom remains. It is this one degree of freedom that corresponds to a very heavy fundamental particle: the Higgs boson H that we observe. ([325]) Note that the VEV of the Higgs doublet field ϕ is not to be confused with the Higgs boson.

It is a subtle concept, so it is worthwhile to reiterate it. In the very early universe, before electroweak symmetry breaking, charged fermions (i.e. electrons and quarks) were massless. There was no electromagnetism yet. Moreover, there were four vector bosons: two massless photons and two massless photon-like particles that carried electric charge. And there was also a complex doublet field, the Higgs field, with some rather weird properties. But as the universe expanded and cooled, the vacuum underwent a phase transition to a lower energy state ([326]). It meant creating a 'sea' of Higgs doublet excitations that became combined with quarks, electrons, and three out of the four vector bosons. So these particles started to interact with the new, lower energy vacuum state in ways they had not done before. This altered their behaviour: with respect to this new vacuum, they now behave as though they were massive particles. Meanwhile, these interactions also altered the Higgs field itself. Two complex fields mean four degrees of freedom, but three of these ended up being tied to the three now massive vector bosons. The remaining degree of freedom is what we know today as the Higgs boson. It is a very massive particle with a very short lifetime, produced in the largest particle accelerator only to decay right away, playing no role of any significance in everyday physics ([T5]).

In technical terms, before the phase transition the electroweak theory has an $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ symmetry with four massless gauge bosons W_μ^1, W_μ^2, W_μ^3 and B_μ. When these bosons interact with the Higgs doublet field, a field that fills the Universe, the symmetry breaks. The Higgs potential gives mass to W_μ^1 and W_μ^2. Consequently, there are two 'new' massive charged bosons $W^+$and W^-, which can be identified as a linear combination of the W_μ^1 and W_μ^2 components of the $\boldsymbol{W}_\mu$ fields:

$$W_\mu^{\pm} = 1/\sqrt{2}(\, W_\mu^1 \mp \mathrm{i} W_\mu^2).$$

---

[324] After rolling down into a point of lower energy, the field ϕ acquires a nonzero positive vacuum expectation value. Physicists discuss the question whether it is the lowest possible energy state of the Higgs. The unpleasant answer is that there might be an even lower energy state (possibly, a state that is not bounded from below at all). This is due to quantum corrections suggesting that in the potential $V(\phi) = \mu^2\phi^\dagger\phi + \lambda(\phi^\dagger\phi)^2$ the term $\lambda(\phi^\dagger\phi)^2$ is replaced by something proportional to $(\lambda^2 - \kappa^2)(\phi^\dagger\phi)^2\log(\phi/\mu)$, where μ is an energy scale (the renormalisation scale) and κ is some constant. If $\lambda < \kappa$, this term is positive when $\phi < \mu$, but becomes negative (and unbounded from below) for $\phi > \mu$. This means that there is a potential barrier beyond it the initially positive energy can 'cascade down' through the negative energy levels, without limit. And no matter how large that potential barrier is, given enough time the probability that it will be breached, if nothing else, by quantum tunneling, will approach unity.
Is this really in our future? Or is there a new lowest energy limit due to some high energy quantum behavior of which we are completely ignorant? Or perhaps this lower energy state does not exist at all? We do not know. But in its simplest form, the Standard Model seems to predict that such a collapse will eventually happen. ([T5])

[325] The Goldstone bosons effectively become the longitudinal parts of the $W^{\pm}$ and Z^0 fields, while the quantised excitations of the fourth Higgs field away from its vacuum value appear physically as neutral spin-0 particles, called Higgs bosons. Note that massless particles (like photon) do not have longitudinal polarization state whereas all massive ones have. Thus when $W^{\pm}$ and Z^0 acquire mass they must also acquire longitudinal polarization (i.e. polarization in the direction of wave propagation). The BEH mechanism precisely furnishes the missing degree of freedom (also called *would-be-Goldstone bosons* in the literature) to make massive bosons out of massless ones ([E1]). For massless vector particles, gauge invariance eliminates one degree of freedom, leaving only two transverse polarization states.

[326] Electroweak symmetry breaking took place when the Universe was already at the respectable age of about a trillionth of a second counting from the presumed initial singularity ([T5]).

The size of the mass of $W^+$and W^- depends on the coupling strength g which is the same for both particles. Things are different for W_μ^3 and B_μ, because they interact with Higgs only in a mixed state. Since there are two possible mixes, one massless and one massive, we get the photon (massless) and the third weak boson Z^0 (massive). The Z^0 and photon field are:

$$Z_\mu^0 = -B_\mu \sin\theta_W + W_\mu^3 \cos\theta_W,$$

$$A_\mu = B_\mu \cos\theta_W + W_\mu^3 \sin\theta_W.$$

The weak mixing angle θ_W has an experimentally determined value of $\sin^2\theta_W \approx 0.23$ ([327]).

So we see that the old W_μ^3-boson gets mixed with the old B_μ-boson. New particles emerge from this mix ([S0]). As a result, there are three 'new' massive bosons: (W^+, W^- and Z^0); the fourth boson γ (photon) is massless – and there is a massive scalar (i.e. spin-0) Higgs boson H left over.

The charged (fundamental) fermions also get mass from a new type of interactions (Yukawa interactions) with the Higgs field. The actual observed mass of these fermions arises as a result of the Yukawa coupling constant that couples the fermions to the Higgs field ([328]). The values of these coupling constants are different for different fermions and span several orders of magnitude. However, not all mass is due to interaction with the Higgs VEV. Neutrinos, as far as we know, get their masses in a way not related to the Higgs at all. In addition, it has been estimated that only approximately 1% of the mass of hadrons is attributable to interaction with the Higgs field. The bulk of their mass is due to the energy associated with strong interactions between quarks and gluons (for further details, please refer to Section 7.5).

In 1971 Gerardus 't Hooft ([329]) showed [H8] that the electroweak theory was renormalisable. Weinberg and Salam had conjectured this, but there was no proof initially. ([330])

The Z^0 boson adds new interactions, ones with neutral currents. The existence of weak neutral currents is a dramatic prediction of the GSW model. The discovery of such interactions was made in 1973 by A. Lagarrigue, P. Musset, D. H. Perkins, A. Rousset and co-workers using the Gargamelle bubble chamber at CERN near Geneva, Switzerland.

And finally, in 2012 a subatomic particle with expected properties of a Higgs boson was discovered by the ATLAS and CMS experiments at the Large Hadron Collider (LHC) at CERN.

The electroweak theory was here to stay.

---

[327] The value of θ_W varies as a function of the momentum transfer, Δq, at which it is measured.

[328] Normally fields only interact via exchange of energy quanta, but the Higgs is different. Field interactions also need to have a non-zero 'coupling constant', a number that indicates how strong their connection is: this is also true of the Higgs couplings to fermions, which are named Yukawa couplings. Now it turns out that since the Higgs field is spinless, its mathematical representation in this Yukawa interaction is just a number. And the combination of the Yukawa coupling and the Higgs VEV looks (mathematically) exactly like a mass term would. So the constant interactions of a massless fermion with the Higgs VEV make it behave exactly as if it has a mass.
Note that Higgs bosons do not permeate the Universe. The Higgs field does. Higgs bosons are excitations (higher-energy states) of the (scalar) Higgs field; they tend to exist very rarely and very briefly, before decaying to other things. It means that particles do not interact with the Higgs field by exchanging Higgs bosons. Rather, particles acquire mass through their interaction with the Higgs VEV, a value that is nonzero even when there are no actual Higgs bosons around. ([T5])

[329] Gerardus (Gerard) 't Hooft (1946 –) is a Dutch theoretical physicist. He shared the 1999 Nobel Prize in Physics with his thesis advisor Veltman.

[330] It was a major breakthrough. "…*the psychological effect of a complete proof of renormalisabiity has been immense*," wrote Veltman some years later. In fact, what 't Hooft had done was demonstrate that Yang–Mills gauge theories in general are renormalisable. Local gauge theories are actually the only class of field theories that can be renormalised. 't Hooft was just 25 years old. Initially, Glashow didn't understand the proof. Of 't Hooft he said: "*Either this guy's a total idiot or he's the biggest genius to hit physics in years*." Weinberg did not believe it, but when he saw that a fellow theorist was taking it seriously he decided to look more closely at 't Hooft's work. He was quickly convinced. ([B1]).

7.5. The strong interaction – QCD

A gauge theory of the strong interaction could not be developed until a fundamental fact about the hadrons (i.e. strongly interacting particles, e.g. protons and neutrons) was understood: they are not fundamental particles. Evidence for the composite nature of hadrons accumulated during the 1960s and 1970s.

Before this discovery, Murray Gell-Mann ([331]) and independently Juval Ne'eman ([332]) proposed in 1961 a phenomenological model for the classification of the hadrons, which Gell-Mann initially named 'Eightfold Way' according to the Noble Eightfold Path in Buddhism, since the number 8 plays a central role in the model ([333]).

Then in 1964, Gell-Mann and independently George Zweig ([334]) proposed a model of hadrons as composite objects. Though based on somewhat different (and much more fragmentary) evidence, their suggestion has turned out to be essentially correct. They proposed that baryons contain three spin-½ constituents called *quarks*, while mesons are quark-antiquark systems.

Quarks ([335]) have different flavours (e.g. up and down) and only carry fractions of an electrical charge, but they always combine in such a way that hadrons have an integer net electrical charge – see Chapter 2.

Any theory of the strong interactions has to explain the peculiar rules of building hadrons out of quarks. The structure of a meson is not too difficult to explain. Since it consists of a quark and antiquark, it is enough to assume that the quarks carry some 'charge' analogous to electric charge and that opposite charges attract. The structure of baryons, however, is more mysterious. To explain how three quarks form a composite object one must assume that three like charges attract ([H9]).

The analogue of electric charge is a new quantum number called *colour* ([336]). The rules for forming hadrons can be expressed by requiring all allowed combinations of quarks to be 'white'

---

[331] Murray Gell-Mann (1929 –2019) was an American physicist who received the 1969 Nobel Prize in Physics.

[332] Yuval Ne'eman (1925 – 2006) was an Israeli theoretical physicist, military scientist, and politician. He had commanded an infantry battalion in the 1948 Arab–Israeli war and served as acting head of the Israeli Secret Service. He had achieved the rank of Colonel in the Israeli Defence Force when he decided to seek an opportunity to study for a doctorate in physics. Moshe Dayan, defence chief of staff, agreed to appoint him as a defence attaché at the Israeli Embassy in London. Dayan figured that Ne'eman could study for his PhD in his spare time. ([B1])

[333] *In December of 1960, after a chance conversation with a Caltech mathematician, Gell-Mann saw how to make the particles fit together beautifully in groups of eight. The Israeli physicist Yuval Ne'eman came up with the same idea at the same time, calling it by its mathematical name, SU(3). Gell-Mann, as usual, picked the name that endured: because he had been reading about Buddhism, he decided to call his classification scheme the Eightfold Way, a mocking reference to the Buddha's eight-step plan for righteous living* ([J2]).

[334] George Zweig (1937 –) is an American physicist. He was trained as a particle physicist under Richard Feynman. Zweig was working as a postdoctoral associate at CERN, and published his model (which he called 'ace model') as a CERN preprint in January 1964. Having subsequently seen Gell-Mann's paper, he moved quickly to elaborate the model, produced a second, 80-page CERN preprint, and submitted this to the prestigious journal *Physical Review*. He was shouted down by his peer reviewers. The paper was never published. When shortly afterwards he sought an appointment at a leading university, one of the faculty members, a respected senior theorist, declared the ace model to be the work of a charlatan. ([B1])

[335] The word 'quark' was coined by Gell-Mann in 1963 taken from „*Finnegans Wake*" by James Joyce:

Three quarks for Muster Mark!
Sure he has not got much of a bark
And sure any he has it's all beside the mark.

Gell-Mann [G2]: "*When I assigned the name 'quark' to the fundamental constituents of the nucleon I had the sound first, without the spelling, which could have been 'kwork.' Then, in one of my occasional perusals of Finnegans Wake, by James Joyce, I came across the word 'quark' (...). I had to find an excuse to pronounce it as 'kwork' (...) I argued, therefore, that perhaps one of the multiple sources of the cry 'Three quarks for Muster Mark' might be 'Three quarts for Muster Mark!' in which case the pronunciation 'kwork' would not be totally unjustified.*"

or colourless. The quarks are assigned the primary colours *red* (r), *green* (g) and *blue* (b). The antiquarks have complementary anti-colours *cyan* ($\bar{r}$), *magenta* ($\bar{g}$) and *yellow* ($\bar{b}$). For example, the proton consists of the following combination: (uud).

Each of the quark flavours comes in all three colours so that the introduction of the colour triples the number of distinct quarks. Colours are charges of the strong interaction, i.e. the strong force acts on the colour charge. It does not differentiate between quark flavours – it is flavour blind. The electromagnetic and weak interactions, on the other hand, are colour blind.

The quark model is complicated by the fact that quarks have eight colour charges, which are non-commutative. Consequently, the effects of these charges do not add in a simple way.

The theory of strong interactions is modelled on quantum electrodynamics QED and is called *quantum chromodynamics* or QCD for short. It is a non-Abelian gauge theory. The gauge symmetry is an invariance with respect to local transformations of quark colour that build a group SU(3), denoted $\mathrm{SU}(3)_c$ and called *colour group* (indicated by the subscript c).

We have three spin-$\frac{1}{2}$ Dirac fields, i.e. the matter field is a 3-component object $\psi = [\psi_r \quad \psi_b \quad \psi_g]^T$ with colour index (r, b, g), and each component is by itself a 4-component Dirac spinor. We ignore here the fact that quarks of different flavours are not identical and do not have the same mass. Thus we should introduce a flavour index f = (d, u, s, c, b, t), and different masses m_f. We will not do that here to keep the notation simple.

To make the Lagrangian density of QCD gauge invariant under local $\mathrm{SU}(3)_c$ transformations, we have to apply the gauge principle. Recall (see Example 7.9) that the algebra $\mathfrak{su}(3)$ has eight generators

$$T_A = \frac{1}{2}\lambda_A,\ \mathrm{A} = 1, 2, ..., 8,$$

defined by the Gell-Mann matrices λ_A. As in other gauge theories, we have to introduce eight 'compensating' gauge vector fields

$$G_\mu^A,\ A = 1, 2, ..., 8,$$

one field for each generator T_A. The covariant derivative D_μ is then defined as

$$D_\mu := \nabla_\mu - \mathrm{i}g_s \sum_{A=1}^{8} T_A\, G_\mu^A = \nabla_\mu - \mathrm{i}g_s G_\mu,$$

where $G_\mu := T_A G_\mu^A := \sum_{A=1}^{8} T_A\, G_\mu^A$, g_s is a 'strong charge' coupling strength and G_μ^A (A = 1, 2, . . . , 8) is an octet of *colour fields*. Colour charges are sources of these fields that give rise to the strong force. The covariant derivative D_μ is now a 3×3 matrix in colour space ([337]).

The quanta of the colour fields are called *gluons*. They are massless spin-1 bosons like photon. Also like photon, gluons are electrically neutral, but they are not colour neutral. Each gluon carries one colour and an anticolour. Gauge invariance requires a single coupling constant g_s, i.e., all eight gluons couple with the same strength to the quarks.

There are nine possible combinations of colour and anticolour, but one of them is equivalent to white and is excluded, leaving eight distinct gluons. Six of them can convert the colour of a quark. These are: $\mathrm{r}\bar{\mathrm{b}}$, $\mathrm{r}\bar{\mathrm{g}}$, $\mathrm{b}\bar{\mathrm{g}}$, $\mathrm{b}\bar{\mathrm{r}}$, $\mathrm{g}\bar{\mathrm{b}}$, $\mathrm{g}\bar{\mathrm{r}}$. For example, red-antigreen ($\mathrm{r}\bar{\mathrm{g}}$) turns a red quark into a green quark. In addition, there are two different gluons that couple to the colour charge on a quark without changing the quark colour:

$$1/\sqrt{2}(\mathrm{r}\bar{\mathrm{r}} - \mathrm{g}\bar{\mathrm{g}}) \text{ and } 1/\sqrt{6}(\mathrm{r}\bar{\mathrm{r}} + \mathrm{g}\bar{\mathrm{g}} - 2\mathrm{b}\bar{\mathrm{b}}).$$

The ninth combination

---

[336] R. Feynman on the designation of this property as 'colour' [F10]: „*The idiot physicists, unable to come up with any wonderful Greek words anymore, call this type of polarization by the unfortunate name of ‚color', [sic!] which has nothing to do with the color in the normal sense.*"

[337] The ∇_μ is multiplied implicitly by the identity 3×3 matrix.

$$1/\sqrt{3}(r\bar{r} + g\bar{g} + b\bar{b})$$

is colourless and does not participate in the strong interaction.

By (7.24) the field strength tensor $\boldsymbol{F}_{\mu\nu}$ is

$$\boldsymbol{F}_{\mu\nu} := \nabla_\mu \boldsymbol{G}_\nu - \nabla_\nu \boldsymbol{G}_\mu + g_s \boldsymbol{G}_\mu \times \boldsymbol{G}_\nu,$$

where $\boldsymbol{G}_\mu := (G_\mu^1, G_\mu^2, ..., G_\mu^8)$. The Lagrangian density for a quark field ψ with the colour degree of freedom (i.e. the wave function ψ with three components) is

$$\mathcal{L}_{QCD} = \bar{\psi}(i\gamma_\mu \nabla^\mu - m)\psi - \frac{1}{4}\boldsymbol{F}_{\mu\nu}\boldsymbol{F}^{\mu\nu} - g_s \bar{\psi}\gamma_\mu G^\mu \psi. \tag{7.41}$$

The first term looks like the free Dirac Lagrangian density $\mathcal{L}_{Dirac}^{free}$, the second term is the kinetic Lagrangian of the gluon fields $\boldsymbol{G}_\mu$ and, finally, the last term is an interaction term. It couples the gluon fields $\boldsymbol{G}_\mu$ with the quark field ψ.

Since the group $SU(3)_c$ is non-Abelian, the field strength tensor $\boldsymbol{F}_{\mu\nu}$ contains a nonlinear term, which means that gluon self-interactions become possible. A mass term is missing in the Lagrangian density $\mathcal{L}_{QCD}$ because the gauge fields, i.e. the gluons, are massless. This is a bit surprising because the strong interaction is extremely short-range. This time the problem cannot be solved by spontaneous symmetry breaking, since the gluons have to be massless.

Notice that quantum electrodynamics and quantum chromodynamics are similar in form: the photon and the gluon are identical in their spin and their lack of mass and electric charge. Yet the interactions of quarks are very different from those of electrons. Both electrons and quarks form composite objects, namely atoms for electrons and hadrons for quarks. However, in contrast with quarks, a small quantity of energy is enough to isolate an electron from an atom.

An isolated quark has never been detected ([338]), no matter how much energy is supplied to ionize a hadron. When hadrons of the highest energies currently available are smashed into each other, what is observed downstream is only lots more hadrons, not fractionally charged quarks ([H9]). This phenomenon may be compared to the division of a magnetic bar into two smaller magnets, as opposed to two monopoles.

However, probes of the internal structure of hadrons show the quarks moving freely as if they were not bound at all. So we are faced with an almost paradoxical situation because we know that the forces are indeed so strong that no one has yet succeeded in separating completely either a quark or a gluon from a hadron so that they emerge as free particles. If you try to separate two quarks, a gluon band forms between them, in which at some point so much energy is stored that one new pair of a quark and the corresponding antiquark is created. This is due to the fact that the potential energy between two quarks increases with their distance until it reaches a certain threshold to produce a new pair of particles. So you cannot isolate a single quark.

Since quarks and gluons carry colour charge, while hadrons do not, this suggests the rule that only 'colourless' states can exist. This phenomenon is known as *colour confinement* or *quark confinement* ([339]). Colour charge generates colour flux lines in a manner analogous to the generation of electric flux lines by electric charge. The energy expended per unit length is a consequence of this process. An explanation for quark confinement is that a colour charge seeks

---

[338] With the exception of the t quark. Its mass is so large that, although it decays weakly, the energy release is so great that its lifetime of $5 \cdot 10^{-25}$s is some two orders of magnitude shorter than typical strong interaction timescales; this means that it decays before any t-carrying hadrons can be formed. So when a t quark is produced, it decays as a free (unbound) particle via the weak interaction. The detection of those decay products allows a direct measurement of the top quarks properties, which is a unique behaviour within the Standard Model.

[339] In contrast to asymptotic freedom, confinement has so far not been deduced from the underlying equations of QCD based on the Lagrangian (7.41). ([E1]) There are strong indications from numerical simulations that it is so, but we do not have a proof. ([J1])

out neutralising partners to form a bound state. The purpose of this process is to contain the flux lines within a microscopic volume, thus minimising energy. ([H13])

Gluons too have not been seen directly in experiments. But if massless particles that so closely resemble the photon existed, they would be easy to detect and they would have been known long ago. On the other hand, giving the gluons a mass via the BEH mechanism would mean that the mass should be large or the gluons would have been produced by now with high-energy accelerators. However, if the mass is large, the range of the strong force becomes too small ([H9]).

The resolution of this quandary lies in the fundamental feature of non-Abelian gauge theories called *asymptotic freedom*, whereby the effective coupling strength becomes progressively smaller at short distances or high energies ([340]). Because the gluons carry colour, they can interact with themselves, like the W's and Z's of the GSW theory. As in that case, these gluonic self-interactions cause the QCD interaction strength to decrease at short distances (or high energies), ultimately tending to zero. This result was first obtained by Politzer (1973), Gross and Wilczek (1973) ([341]) and 't Hooft. 't Hooft's result, announced at a conference in Marseilles in 1972, was not published ([342]) ([A1], [W11]).

The result of Politzer and of Gross and Wilczek led rapidly to the general acceptance of QCD as the theory of strong interactions, a conclusion reinforced by the demonstration by Coleman ([343]) and Gross (1973) that no theory without Yang-Mills fields possessed the property of asymptotic freedom.

The short range of the strong force is explained by the fact that the potential of the QCD has special properties. The range of the strong force is, kind of paradoxically, limited by the fact that it becomes stronger with distance. The coupling strength g_s increases with distance – at large distances (≡ low energies) g_s is large. Consequently, quarks and gluons are confined inside hadrons, i.e. there are no free quarks/gluons that would be far apart.

But why is QCD asymptotically free, in contrast to QED? Unfortunately, a proper understanding of how it all works necessitates a considerable detour into the physics of

---

[340] This is due to the fact that the strong coupling constant g_s depends on the energy scale of the interaction and decreases with higher energy. More generally, it is a consequence of the renormalisation procedure that the physical (i.e. renormalised) coupling constants are not at all constant. Rather, their value depends on the energy scale at which they are measured. If the coupling goes to zero in the high-energy limit the theory is called asymptotically free. The theory of strong interactions, QCD, turns out to be an asymptotically free theory. That is why quarks are described as *asymptotically free* particles. ([M1])

[341] In late 1972 Princeton theorist David Gross had set out to show that asymptotic freedom was simply impossible in a quantum field theory. With the help of his student Frank Wilczek he managed instead to prove precisely the opposite. Quantum field theories based on local gauge symmetries can accommodate asymptotic freedom. A young Harvard graduate student called David Politzer independently made the same discovery. Their papers were published in the June 1973 issue of *Physical Review Letters*. ([B1])

David Jonathan Gross (1941 –) is an American theoretical physicist and Professor of Theoretical Physics at the Kavli Institute for Theoretical Physics of the University of California, Santa Barbara (UCSB),
Hugh David Politzer (1949 –) is an American theoretical physicist and Professor of Theoretical Physics at the California Institute of Technology.
Frank Anthony Wilczek (1951 –) is an American theoretical physicist and Professor of Physics at the Massachusetts Institute of Technology (MIT).
Gross, Politzer and Wilczek shared the Nobel Prize in Physics in 2004.

[342] 't Hooft had already concluded that Yang–Mills gauge theories could show this counter-intuitive behaviour, but he was busy working on renormalization at this time and did not follow it up. ([B1])

[343] Sidney Richard Coleman (1937 – 2007) was an American theoretical physicist who studied under Murray Gell-Mann.

renormalisation ([344]), but it would go beyond the scope of this paper.

As previously mentioned in Section 7.4, quarks acquire mass through interactions with the Higgs field. What we observe in the laboratory, however, are not quarks, but hadrons composed of quarks. The masses of hadrons have no direct relation to the quark masses, and have little to do with the Higgs field. Only roughly 1% of the mass of hadrons comes from interaction with the vacuum expectation value of the Higgs field. To illustrate this point, consider the theoretical masses of the light quarks u and d, which are respectively 2 and 4.8 MeV. However, the proton, composed of uud, has a mass of 938 MeV. This demonstrates that the quark masses are negligible in comparison to the proton mass. ([H13])

The majority of the proton mass comes from the energy associated with the strong interactions between quarks and gluons, rather than from the masses of the quarks themselves. The spontaneous breaking of chiral symmetry is a key mechanism in this process, as it leads to the generation of the effective quark masses and the binding energy that holds the proton together. In the chiral limit, where the masses of the up and down quarks are set to zero, QCD exhibits chiral symmetry. This symmetry signifies that the left-chiral and right-chiral components of the quark fields are to be regarded as distinct entities – see (5.28').

In the QCD vacuum, this chiral symmetry is spontaneously broken. This means that the ground state of the theory does not respect the symmetry, even though the underlying equations do. This breaking of symmetry gives rise to the masses of hadrons (such as protons and neutrons) through the interactions of quarks and gluons. The quark condensate interacts with the quark fields, leading to the generation of *effective masses* for the quarks. These effective masses are much larger than the masses of the quarks due to their interaction with the Higgs field and, as result, contribute significantly to the mass of the proton. ([H13])

In summary, QCD is a renormalisable relativistic quantum field theory based on the $SU(3)_c$ gauge symmetry. The technical problems involving QCD calculations cause that the agreement with experiments is much less impressive than is the case for QED, but there are many reasons to suppose that QCD is the theory that describes hadronic physics ([W0]).

We close this section by briefly discussing the requirement of renormalisability for quantum field theories. The physical motivation comes from the idea that QFTs involve an implicit maximum energy beyond which they cannot be applied. In other words they are valid up to energies below a certain scale Λ, which then represents an ultraviolet (UV) cut-off of a field theory. In classical physics it is not important to specify such a cut-off carefully. In a quantum

---

[344] The early formulators of quantum field theories were, as a rule, dissatisfied with the renormalisation techniques. It seemed illegitimate to do something tantamount to subtracting infinities from infinities to get finite answers. F. Dyson argued that these infinities are of a basic nature and cannot be eliminated by any formal mathematical procedures, such as the renormalisation method. Dirac (1942) proposed to abandon unitarity, but physical consequences seemed hardly acceptable. Wheeler (1937) and Heisenberg (1943) proposed to completely abandon QFT in favour of a theory of physical observables (scattering data) – the so-called S-matrix theory, a somewhat desperate idea that nevertheless became very popular in the 1960's. ([Z3])
Another important critic was Feynman. Despite his crucial role in the development of quantum electrodynamics QED, he wrote the following in 1985 ([F10]):
"The shell game that we play is technically called 'renormalization'. But no matter how clever the word, it is still what I would call a dippy process! Having to resort to such hocus-pocus has prevented us from proving that the theory of quantum electrodynamics is mathematically self-consistent. It's surprising that the theory still hasn't been proved self-consistent one way or the other by now; I suspect that renormalization is not mathematically legitimate."

Beginning in the 1970s, however, inspired by work on the renormalisation group and effective field theory, attitudes began to change, especially among younger theorists. Justification for normalisation came decades later from a seemingly unrelated branch of physics. Researchers studying magnetisation discovered that renormalisation wasn't about infinities at all. Instead, it spoke to the universe's separation into domains of independent sizes, a perspective that guides many corners of physics today.
"*Renormalization helps us simplify the problem*," said Nathan Seiberg, a theoretical physicist at the Institute for Advanced Study in Princeton, New Jersey. But "*it also hides what happens at short distances. You can't have it both ways*." ([W11], [W16]).

theory, however, since all states can contribute to any given process as intermediate (or 'virtual') particles, any quantum calculation will depend explicitly on the cut-off scale Λ. For example, the quantum electrodynamics of electrons and photons is only physically correct up to an energy of twice the mass of the lightest particle that is heavier than the electron: $\Lambda = 2m_{\mu}$, i.e. twice the muon mass. At energies higher than this, muons can no longer be neglected, since they can be pair-produced in the quantum process under consideration. The correct theory for physics at energies above Λ becomes the quantum electrodynamics of photons, electrons and muons ([345]). This theory is in turn only valid up to the next threshold, and so on ([B11]).

If detailed knowledge of physics at the Λ scale is necessary in order to calculate probability amplitudes for processes at energies lower than Λ, then the theory is called *non-renormalisable* In *renormalisable* theories, on the other hand, Λ only appears in physical predictions (for large Λ) through a small number of parameters, such as the masses and charges, whose values have to be determined experimentally. All other processes may then be computed in terms of these parameters and definite predictions are possible. For example, in QED there are only two such parameters: the mass and charge of electron ([B11]).

This physical picture implies that renormalisability is the minimal criterion for a theory which aims to describe all of the physics appropriate to any given scale. Demanding renormalisability for the SM then amounts to the assumption that no unknown particles or interactions are required to understand present experiments (see [B11] for more details).

8. Postlude

Today, all fundamental interactions are described by gauge theories. As a matter of fact, this is the only way that was found to describe the forces in nature in a mathematically consistent way. Without gauge theory the Standard Model (SM) of particle physics cannot be formulated and it is basically impossible to fully understand the role of the Higgs field without some understanding of the role of gauge symmetries. The key feature of gauge theories is the concept of a local symmetry. With this we mean that the mathematical transformation that defines the symmetry may be applied differently in different points in space. ([B3])

The Standard Model of the subatomic particles, developed in the 1960s and 1970s, has stood for more than 40 years as 'the' theory of particle physics, passing numerous stringent tests. While many physicists believe that the SM is not a complete description of particle physics, it is expected to be, at worst, incomplete rather than wrong – thus the SM is at worst a subset of the true theory of particle physics ([B11]).

The Standard Model provides a relatively simple picture of quarks and leptons and their non-gravitational interactions. The quark colour triplets are the basic source particles of the gluon fields in QCD, and they bind together to make hadrons. The weak interactions involve quark and lepton doublets. These are sources for the $W^{\pm}$ and Z^{0} fields. Charged fermions (quarks and leptons) are sources for the photon field. All the mediating force quanta have spin 1. The weak and strong force fields are generalizations of electromagnetism; all three are examples of gauge theories but realized in subtly different ways ([A1]).

The Standard Model is a gauge theory based on the three-component Lie group $SU(3)_c \times SU(2)_L \times U(1)_Y$. It offers a seemingly correct and complete description of virtually all fundamental particle phenomena. Its $SU(3)_c$ component yields the unbroken gauge theory of quantum chromodynamics QCD, which ensures quark confinement and underlies nuclear forces. The $SU(2)_L \times U(1)_Y$ is spontaneously broken. Its broken symmetries yield the massive bosons to

[345] Notice, however, that muons would be stable particles, since QED cannot account for the observed muon decay. That decay involves weak interactions about which QED knows nothing. To get the quantum electrodynamics of photons, electrons and muons, it is necessary to go beyond QED and use the electroweak theory. ([T5])

weak interactions; its unbroken $U(1)_Y$ subgroup yields quantum electrodynamics QED ([W0]). The Lagrangian density of the Standard Model can be written as

$$\mathcal{L}_{SM} = \mathcal{L}_{YM} + \mathcal{L}_{Dirac} + \mathcal{L}_{Higgs} + \mathcal{L}_{Yukawa}$$

where

$\mathcal{L}_{YM}$ – kinetic part of the gauge fields
$\mathcal{L}_{Dirac}$ – Dirac fermions
$\mathcal{L}_{Higgs}$ – Higgs dynamics and EWSB (Electroweak Symmetry Breaking)
$\mathcal{L}_{Yukawa}$ – Yukawa sector (interactions between the Higgs doublet and fermions).

Using three symmetry groups to describe three different interactions (weak, strong, and electromagnetic), the Standard Model is *not* a unified theory. A more ambitious theory would embed it within a larger one-component group, what mathematicians call a simple group. Several so-called *Grand Unified Theories* (GUTs) have been proposed, but none has yet proven empirically successful ([W0]).

The Standard Model is not yet a 'theory of everything'. It does not account for the force of gravity. In recent years physicists have developed new theories which attempt to unify the fundamental forces, including gravity. These are theories such as superstrings and *Loop Quantum Gravity* ([S9a]). Despite the efforts of hundreds of theorists engaged on these projects, these new theories remain speculative and have little or no supporting evidence from experiment ([B1]).

For the time being, and despite flaws that have been acknowledged since its inception in the late 1970s, the Standard Model is still where most of the real action is ([B1]).

Acknowledgements

The quality of this manuscript was substantially improved over that of early versions through the diligence of Dr H. Kunde (Rostock) who took an interest in what I was doing, and carefully read drafts of the manuscript.

Professor S. Dolecki (Université de Bourgogne et Franche Comté, Dijon) provided a full list of write errors, excellent suggestions for improvements, and valuable comments on structuring this manuscript.

Thanks go also to Professor M. Bulenda (OTH Regensburg) who read the manuscript, gave me a feedback and encouraged me to publish it.

References

[A0] Y. Aharonov and D. Bohm, *Significance of Electromagnetic Potentials in the Quantum Theory*, Phys. Rev. 115 (3), 1959, 485–491.

[A1] I. Aitchison and A. Hey, *Gauge theories in particle physics: a practical introduction*, Vol. 1 and 2, CRC Press, 2013.

[A2] APS News, *January 1925: Wolfgang Pauli announces the exclusion principle*, 16(1), January 2007.

[A3] J. Atkins, *The Free Klein Gordon Field Theory,* University of Rochester, 2018.

[B0] J. Baez and J. Huerta, *The Algebra of Grand Unified Theories*, Bull. Amer. Math. Soc. 47, 2010, 483–552.

[B1] J. Baggott, *Higgs -The Invention and Discovery of the 'God Particle'*, Oxford University Press, 2012.

[B2] N. Beisert, *Symmetries in Physics*, ETH Zurich, 2015.

[B3] N.-E. Bomark, *Introduction Teaching gauge theory to first year students,* arXiv:2009.02162v2

[B4] M. Bonitz, *Introduction to Quantum Field Theory and Quantum Statistics,* Kiel University, April 22, 2021.

[B5] V. Bouchard, *Group Theory in Physics: Lecture Notes*, University of Alberta, 2020.

[B6] S. Boyd, *The weak interaction*, University of Warwick, 2020.

[B7] K. Brading, E. Castellani (editors), *Symmetries in physics: Philosophical reflections*, Cambridge University Press, 2003.

[B8] M. Breinig, *Quantum mechanics,* University of Tennessee, Knoxville.

[B9] Britannica, https://www.britannica.com/

[B10] A. Buckley, Various blog posts on https://www.quora.com.

[B11] C. P. Burgess and Guy D. Moore, *The Standard Model: A Primer*, Cambridge University Press, 2007.

[C1] S.M. Carroll, *Lecture Notes on General Relativity*, University of California, Santa Barbara, 1997.

[C2] B. Cassen and E. U. Condon, *On Nuclear Forces*, Phys. Rev., 50(9), 1936, 846–849.

[C3] M. Chaves, *An introduction to generalized Yang-Mills theories*, Hadronic J. Suppl., 17, 2002, 3 – 51.

[C4] CMS Collaboration, *Measurement of the W boson helicity fractions in the decays of top quark pairs to lepton+jets final states produced in pp collisions at* $\sqrt{s} = 8$ TeV, https://doi.org/10.1016/j.physletb.2016.10.007

[C5] E.D. Commins, *Electron Spin and Its History*, Annu. Rev. Nucl. Part. Sci., 62, 2012,133–157.

[C6] Contemporary Physics Education Project, https://particleadventure.org/

[C7] R.P. Crease, *From wrong to right*, Physics World, October 2015.

[D1] Dictionary.com, https://www.dictionary.com/

[D2] P.A.M. Dirac, *The Principles of Quantum Mechanics*, Oxford University Press, 1958.

[D3] *Dirac Equation and Spinors*, http://physics.gu.se/~tfkhj/TOPO/DiracEquation.pdf

[D4] DoITPoMS, *What is a tensor*, University of Cambridge, https://www.doitpoms.ac.uk/tlplib/tensors/what_is_tensor.php

[E1] G. Ecker, *Particles, Fields, Quanta. From Quantum Mechanics to the Standard Model of Particle Physics*, Springer, 2019.

[E2] *Einstein summation convention and δ-functions,* https://www.dr-qubit.org/teaching/summation_delta.pdf

[E3] A. Einstein, *Uber einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristishen Gesichtspunkt*, Ann. Phys. 17, 1905, 132–148.

[E4] A. Einstein, *Die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen*, Ann. Phys. 17, 1905, 549–560.

[E5] A. Einstein, *Zur Elektrodynamic bewegter Körper*, Ann. Phys. 17, 1905, 891–921.

[E6] A. Einstein, *Ist die Trägheit eines Körpers von seinem Energienhalt abhängig?* Ann. Phys. 18, 1905, 639-641.

[F1] G. Farmelo, *The Strangest Man. The Hidden Life of Paul Dirac, Quantum Genius*, Faber and Faber Ltd., 2009.

[F2] R. Feynman, *The Feynman Lectures on Physics*, https://www.feynmanlectures.caltech.edu/II_01.html

[F3] R. Feynman, *The Feynman Lectures on Physics*, https://www.feynmanlectures.caltech.edu/II_18.html

[F4] R. Feynman, *The Feynman Lectures on Physics*, https://www.feynmanlectures.caltech.edu/II_19.html

[F5] R. Feynman, *The Feynman Lectures on Physics*, https://www.feynmanlectures.caltech.edu/II_25.html

[F6] R. Feynman, *The Feynman Lectures on Physics*, https://www.feynmanlectures.caltech.edu/II_26.html

[F7] R. Feynman, *Symmetries in Elementary Particle Physics*, 1964 Int. School of Physics 'Ettore Majorana' ed. A Zichichi (New York: Academic Press).

[F8] R. Feynman, *Feynman Lectures On Gravitation*, Lectures given at Caltech during the 1962-63 academic year.

[F9] R. Feynman, *The character of physical law*, Lectures in 1964 at Cornell University, as part of the Messenger Lectures series, Penguin Books, 1992.

[F10] R. Feynman, *QED: The Strange Theory of Light and Matter*, Princeton University Press, 2006.

[F11] R. Feynman, *What is Science?*, The Physics Teacher 7 (6), 1969, 313 pp.

[F12] R. Feynman, *Negative probability*, In Basil J. Hiley & D. Peat (eds.), Quantum Implications: Essays in Honour of David Bohm, 1987, 235–248.

[F13] F. Finster, *The Principle of the Fermionic Projector,* https://arxiv.org/pdf/hep-th/0001048.pdf

[F14] C. Foudas, *Helicity and Chirality*, Imperial College London, 2009.

[F15] E. Fradkin, *Classical Symmetries and Conservation Laws*, http://eduardo.physics.illinois.edu/phys582/582-chapter3-edited.pdf

[F16] M. Freiberger, *Schrödinger's equation — what does it mean*?, plus.math.org , August 2, 2012.

[G1] M. Gell-Mann, *A Schematic Model of Baryons and Mesons*, Physics Letters, 8, 1964, 214–215.

[G2] M. Gell-Mann, *The Quark And The Jaguar: Adventures in the Simple and the Complex,* W H Freeman & Co, 1995.

[G3] S. Glashow, *The Yang-Mills Model*, Inference: International Review of Science, 5(2), 2020.

[G4] D.L. Goldstein, *Richard P. Feynman, Teacher*, Physics Today, 42(2), 1989, 70–75.

[G5] J. Goldstone, *Field theories with «Superconductor» solutions*, Nuovo Cimento, 19, 1961, 154–164.

[G6] D. Griffiths, *Introduction to Elementary Particles,* VILLEY-VCH Verlag, 2008.

[G7] B. Gripaios, *Gauge Field Theory,* University of Cambridge, 2016.

[H0] B. Hall, *Lie Groups, Lie Algebras, and Representations,* Springer, 2015.

[H1] T. Hällgren, *Yang-Mills Theory,* KTH Royal Institute of Technology, Stockholm, Sweden.

[H2] M.J.D. Hamilton, *Mathematical Gauge Theory,* Springer, 2017.

[H3] W. Heisenberg, *Über den Bau der Atomkerne*. I., Zeitschr. f. Phys., 77, 1932, 1–11.

[H4] T.M. Helliwell, *Special Relativity*, University Science Books, 2010.

[H5] J. Hershey, *Faithful Science*, http://www.faithfulscience.com

[H6] P.W. Higgs, *Broken symmetries, massless particles and gauge fields*, Physics Letters, 12(2), 1964, 132–133.

[H7] A. Hobson, *There are no particles, there are only fields*, American Journal of Physics, (81), 2013, 211–223.

[H8] G. 't Hooft, *Renormalization of Massless Yang-Mills Fields*, Nucl. Phys., B33, 1971, 173–199.

[H9] G. 't Hooft, *Gauge Theories of the Forces between the Elementary Particles*, Scientific American, 242(6), 1980, 104–138.

[H10] G. 't Hooft, *Gauge theories*, Scholarpedia, 3(12), 2008. doi:10.4249/scholarpedia.7443

[H11] J.P. Hosemann, *Auf Dem Weg Zur Erklärung der Welt: Meilensteine der Physik und Astrophysik*, Logos Berlin, 2014.

[H12] S. Hossenfelder, *Lost in Math: How beauty leads physics astray*, Basic Books, 2018.

[H13] K. Huang, *Fundamental Forces of Nature: The Story of Gauge Fields*, Wold Scientific, 2007.

[H14] K. Huang, *Quarks, Leptons & Gauge Fields*, World Scientific, Singapore, 1992.

[H15] Ch. Hudgins, *Understanding Noether's Theorem with symplectic geometry,* Semantic Scholar, 2017.

[H16] J. Huerta, *Isospin and SU(2),* https://math.ucr.edu/~huerta/guts/node4.html

[H17] C. Hughes, *A brief discussion on representations*, https://studylib.net/doc/18327185/a-brief-discussion-on-representations

[I1] M. Ibe, A. Kusenko, T. Yanagida, *Why three generations?*, Physics Letters, B758, 2016, 365–369.

[I2] C. Itzikson and J.-B. Zuber, *Quantum Field Theory*, McGraw-Hill Inc., 1980.

[I3] I.P. Ivanov, *Building and testing models with extended Higgs sectors*, Prog. Part. Nucl. Phys., 95, 2017, 160-208.

[J1] A. Jaffe and E. Witten, *Quantum Yang-Mills theory*, ResearchGate, 2009.

[J2] G. Johnson, *The Jaguar and the Fox*, 2000, https://www.theatlantic.com/magazine/archive/2000/07/the-jaguar-and-the-fox/378264/

[J3] H.F. Jones, *Groups, Representations and Physics*, Taylor & Francis Group, 1998.

[K0] A. Khrennikov, *Interpretations of Probability*, Walter de Gruyter, 2009.

[K1] Y.S. Kim, *Eugene Paul Wigner: Poincaré Group after Einstein*, https://ysfine.com/wigner/wigpoinc.html

[K2] H. Kleinert, *Weak interactions*, Freie Universität Berlin, 2016.

[K3] A. Kojevnikov, *Dirac's Quantum Electrodynamics*, The University of British Columbia, Einstein Studies (2002) 10, 229-259.

[K4] L.M. Krauss, *Quantum Man: Richard Feynman's Life in Science*, W.W. Norton, 2011.

[K5] W. Kuehn, *Basic topics in Flavour Physics,* 54. International Winter Meeting on Nuclear Physics, 2016.

[L1] T. Lancaster and S. Blundell, *Quantum Field Theory for the Gifted Amateur*, Oxford University Press, 2014.

[M0] A. Maas, *Global and local symmetries,* axelmaas.blogspot.com, August 4, 2010.

[M1] M. Maggiore, *A Modern Introduction to Quantum Field Theory,* Oxford University Press, 2005.

[M2] N. Manton, *Symmetries, Fields and Particles*, University of Cambridge, September 23, 2014.

[M3] MediaWiki, https://www.mediawiki.org/

[M4] N. Miller, *Representation Theory and Quantum Mechanics*, Harvard University, 2018.

[M5] R. Mills, *Space*, *Time and Quanta: An Introduction to Contemporary Physics*, W.H. Freeman and Company, 1994.

[M6] A. Mitov, *Gauge Field Theory*, University of Cambridge, 2021.

[M7] A. Mitov, *The weak interaction and V-A*, University of Cambridge, 2018.

[M8] L. Molt, Blog posts on https://www.quora.com.

[M9] S. Morrison, *Connections on principal fibre bundles*, 2000, https://tqft.net/papers/ConnectionsAndBundles.pdf

[M10] C. Moskowitz, *Q&A: Lawrence Krauss on The Greatest Story Ever Told*, Scientific American, March 21, 2017.

[N1] K. Nguyen, *The Higgs Mechanism*, Ludwig-Maximilians-Universität München, 2009, https://www.theorie.physik.uni-muenchen.de/lsfrey/teaching/archiv/sose_09/rng/higgs_mechanism.pdf

[N2] M. Nielsen, *An introduction to Yang-Mills theory*, https://michaelnielsen.org/blog/yang_mills.pdf

[O1] L. O'Raifertaigh, *The Dawning of Gauge Theory*, Princeton University Press, 1997.

[O2] L. O'Raifertaigh, *The Evolution of the Gauge Principle*, https://dair.dias.ie/669/1/DIAS-STP-96-19.pdf

[O3] L. O'Raifertaigh, N. Straumann, *Early History of Gauge Theories and Kaluza-Klein Theories, with a Glance at Recent Developments*, https://arxiv.org/pdf/hep-ph/9810524.pdf

[O4] N. T. Osakabe et al., *Experimental confirmation of Aharonov-Bohm effect using a toroidal magnetic field confined by a superconductor*, Phys. Rev. A 34(2), 815 (1986).

[P1] A. Pais, *The Geniuses of Science. A Portrait Gallery*, Oxford University Press, 2000.

[P2] W. Pauli, *Relativitätstheorie*. Encyklopädie der Mathematischen Wissenschaften 5.3, Leipzig: Teubner, 1921, 539-775.

[P3] W. Pauli, *Theory of Relativity*. Pergamon Press, New York, 1958.

[P4] W. Pauli, https://microboone-docdb.fnal.gov/cgi-bin/RetrieveFile?docid=953;filename=pauli%20letter1930.pdf

[P5] R. Penrose, *The Road to Reality - A complete guide to the laws of the Universe*, Vintage 2007.

[P6] physics.info, *The Standard Model*, https://physics.info/standard/

[P7] Ch. Pope, *611 Electromagnetic Theory II*, Texas A&M University, http://physikmethoden.weebly.com

[P8] T. Preis, *Quantum Field Theory,* https://thpreis.github.io/files/QFTnotes.pdf

[P9] Postcard from Einstein to Weyl (April 15, 1918), Archiv der ETH-Zürich, https://ethz.ch/content/dam/ethz/associates/ethlibrary-dam/documents/Standorteundmedien/Plattformen/EinsteinOnline/Die-Berliner-Zeit/Dokumente-zur-Zeit-in-Berlin/1918-04-15-Postk-Einstein-an-Weyl-Hs_91_541.pdf

[Q1] A. Quandt, *Top Quark Physics at Hadron Colliders*, in Advances in the Physics of Particles and Nuclei (28), Springer, 2007.

[Q2] C. Quigley, *On the Origins of Gauge Theory*, 2003, http://www.math.toronto.edu/~colliand/426_03/Papers03/C_Quigley.pdf

[R1] D.V. Redzic, *Are Maxwell's equations Lorentz-covariant?,* https://arxiv.org/pdf/1605.05358.pdf

[R2] *Reading Feynman*, https://readingfeynman.org/tag/probability-amplitudes/

[R3] V. Rubakov, *Classical Theory of Gauge Fields*, Princeton University Press, 2002.

[R4] Rukhsan-Ul-Haq, *Geometry of Spin: Clifford Algebraic Approach*, RESONANCE, December 2016.

[S0] W. Schmitz, *Particles, Fields and Forces - A Conceptual Guide to Quantum Field Theory and the Standard Model*, Springer, 2019.

[S1] M.D. Schwartz, *Quantum Field Theory and the Standard Model*, Cambridge University Press, 2014.

[S2] J. Schwichtenberg, *No-nonsense quantum field theory*, No-Nonsense Books, 2020.

[S2a] J. Schwichtenberg, *No-nonsense quantum mechanics*, No-Nonsense Books, 2020.

[S3] J. Schwichtenberg, *Physics from Symmetry*, Springer, 2018.

[S3a] J. Schwichtenberg, *Demystifying Gauge Symmetry*, arXiv:1901.10420v1

[S4] J. Schwichtenberg, *Short Introduction to and Motivation for Representation Theory*, http://jakobschwichtenberg.com/short-introduction-motivation-representation-theory/

[S5] J. Schwichtenberg, *What's so special about the adjoint representation of a Lie group?*, http://jakobschwichtenberg.com/adjoint-representation/

[S6] M. Sener, K. Schulten, *Symmetries in Physics: Isospin and the Eightfold Way*, University of Illinois, 2000.

[S7] R. Shaw, *Ph.D. Thesis*, Cambridge University (Sept. 1955).

[S8] M. Shifman, *Quantum Field Theory II,* World Scientific Publishing, 2019.

[S9] L. Smolin, *Einstein's Unfinished Revolution: The Search for What Lies Beyond the Quantum*, Penguin LCC US, 2019.

[S9a] L. Smolin, *Three Roads to Quantum Gravity*, Basic Books, 2001.

[S10] F.M. Springer, *Symmetries of the Standard Model,* University of Amsterdam, June 26, 2017.

[S11] M. Srednicki, *Quantum Field Theory*, University of California, Santa Barbara, 2006, https://web.physics.ucsb.edu/~mark/ms-qft-DRAFT.pdf

[S12] A.M. Steane, *An introduction to spinors*, 2013, https://arxiv.org/abs/1312.3824

[S13] N. Straumann, *Early History of Gauge Theories and Weak Interactions*, Invited talk at the PSI Summer School on Physics with Neutrinos, Zuoz, Switzerland, August 4–10, 1996.

[S14] N. Straumann, *Gauge Principle and QED*, Invited talk at PHOTON2005, The Photon: Its First Hundred Years and the Future, 31.8-04.09, 2005, Warsaw.

[S15] N. Straumann, *Hermann Weyl and the early history of gauge theories*, in "Symmetries in Algebra and Number Theory", contributions to "On the Legacy of Hermann Weyl", p.173, Universitätsverlag Göttingen, 2009.

[S16] L. Susskind and A. Friedman, *Quantum Mechanics - The Theoretical Minimum*, Penguin Books, 2015.

[S17] L. Susskind and A. Friedman, *Special Relativity and Classical Field Theory - The Theoretical Minimum*, Penguin Books, 2018.

[T1] T. Teubner, *The Standard Model*, University of Liverpool, 2008.

[T2] The Sci_Co_path, *Nature, Symmetry and Physical laws*, July 6, 2019.

[T3] M.A. Thomson, *Handout 13: Electroweak Unification and the W and Z Bosons*, University of Cambridge, 2011.

[T4] D. Tong, *Gauge Theory*, University of Cambridge, 2018.

[T5] V. Toth, Various blog posts on www.quora.com.

[V1] M.B. Valente, *The Dirac equation and its interpretations*, https://www.researchgate.net/publication/340664785

[W0] S. Webb, *Out of this world: Colliding Universes, Branes, Strings, and Other Wild Ideas of Modern Physics*, Praxis Publishing Ltd., 2004.

[W1] S. Weinberg, *The Quantum Theory of Fields. vol. 1: Foundations,* Cambridge University Press, 1995.

[W2] J.D. Wells, *Lectures on Standard Model Particle Physics*, July 4-10, 2013, CERN Summer Student Lecture Programme.

[W3] Ch. Wendl, *Lecture Notes on Bundles and Connections, Appendix A: Multilinear algebra and index notation,* Humboldt-Universität zu Berlin, 2008.

[W4] S. West, *The Standard Model*, University of London, 2010–2014.

[W5] H. Weyl, *Gravitation und Elektrizität*. Sitzungsberichte der Akademie der Wissenschaften Berlin, 1918, 465–480.

[W6] H. Weyl, *Eine neue Erweiterung der Relativitätstheorie*, Ann. Physik, 59 (10), 1919, 101–133.

[W7] H. Weyl, *Elektron und Gravitation*, Z. Phys., 56, 1929, 330–352.

[W8] N. Wheeler, *Classical Gauge Fields*, Reed College Portland, https://www.reed.edu/physics/faculty/wheeler/documents/Classical%20Field%20Theory/Class%20Notes/Field%20Theory%20Chapter%203.pdf

[W9] E. Wigner, *On the Consequences of the Symmetry of the Nuclear Hamiltonian on the Spectroscopy of Nuclei*, Phys. Rev., 51(2), 1937, 106–119.

[W10] E. Wigner, *On unitary representations of the inhomogeneous Lorentz group*, Ann. of Math., 40(1), 1939.

[W11] Wikipedia, https://en.wikipedia.org/

[W12] B. de Wit, *Introduction to Gauge Theories and the Standard Model*, Institute for Theoretical Physics, Utrecht, NL, 1995.

[W13] P. Woit, *Not even wrong*, Basic Books, 2006.

[W14] P. Woit, *Quantum Theory, Groups and Representations*, Springer, 2017.

[W15] P. Woit, *Topics in Representation Theory: The Adjoint Representation*, Department of Mathematics, Columbia University.

[W16] Ch. Wood, *How Mathematical 'Hocus-Pocus' Saved Particle Physics*, Quanta Magazine, September 17, 2020.

[Y1] C.N. Yang, R.L. Mills, *Conservation of Isotopic Spin and Isotopic Gauge Invariance*, Physical Review, 96 (1), 1954.

[Y2] C.N. Yang, *Selected Papers 1945-1980 With Commentary*, Freeman and Co., San Francisco 1983.

[Y3] C.N. Yang, in *50 Years of Yang–Mills Theory*, ed. G. 't Hooft, World Scientific, Singapore, 2005.

[Y4] H. Yukawa, *On the Interaction of Elementary Particles*, Proc. phys. math. Soc. Japan, 17(48), 1935.

[Z1] A. Zee, *Group Theory in a Nutshell for Physicists*, Princeton University Press, 2016.

[Z2] A. Zee, *Fearful Symmetry: the search for beauty in modern physics*, Princeton University Press, 2016.

[Z3] J. Zinn-Justin, http://users.physik.fu-berlin.de/~pelster/Seminar6/zinn-justin-slides1.pdf

[Z4] G. Zweig, *An SU(3) Model for Strong Interaction Symmetry and Its Breaking I+II*. 1964, CERN Preprint CERN-TH-401.

[Z5] B. Zwiebach, *Spin one half, bras, kets, and operators*, MIT, 2013.

[Z6] L. Zyga, *On the origins of the Schrodinger equation*, Phys.org, 2013.