Modern chapter of theoretical particle physics

Background material
An introduction to the foundations of modern physics

SS 2021 at the KFU Graz

Axel Maas

Contents

Chapter 1

Fundamentals of physics

1.1 Introduction

Our current understanding of the fundamental laws of physics consists out of two parts. The first is particle physics. Its theory and details are quite involved, and it is a genuine quantum theory. In its current form it has evolved in the 1970ies. The second part is the non-quantum theory of general relativity. It is much older, from the 1910ies, and, in a sense, much more simple, as it does not require too much about the matter content. While both theories have been extremely successful, explaining many physical phenomena, partly to extreme precision, the situation is far from satisfactory.

The most obvious problem is that these are two disjoint theories. However, entities like electrons exist in both, but behave fundamentally different, and their description cannot be joined. This will be elaborated more later, but the main problem is that one is a quantum theory and one is not.

Beyond that, there are also several other hints, mostly from astrophysics and mathematical consistency, that even a joint version of these theories may not yet be the answer. Among these are questions pertaining to dark matter and the origin of the universe or why there was a big bang. However, it has become more and more clear that any solution of these problems will likely not emerge from either of these theories alone, but will require a common foundation. The search for this common foundation of physics is therefore naturally the most fundamental goal.

The primary aim of this material is to understand the elements of both theories, and the challenges and attempts in joing them. This will naturally cover more of particle physics than gravity, because the larger complexity of it. Though this is not necessarily mathematical complexity, but this is also a question of experience and perspective, it certainly is with respect to the number of phenomena involved.

1.2 Natural units

Before setting out in this, it is necessary to talk about units.

The usual units of the SI system, like meters or grams, are particularly unsuited for particle physics or gravity. The elementary particles live on very short time scales, move at high speeds, and have tiny masses. E. g. typical energy and distance scales are

$$1 \text{ eV} = 1.60 \times 10^{-19} \text{ J} \quad (1.1)$$

$$1 \text{ fm} = 10^{-15} \text{ fm}, \quad (1.2)$$

which therefore have earned their own name, electron volts (eV), and Fermi (being identical to a femtometer). Typical energy scales are actually rather 10^9 eV, a GeV.

Furthermore, in particle physics the system of units is further modified. Since units are man-made, they are not necessary, and the only thing really useful is a single scale to characterize everything[1]. This is achieved in the system of natural units, which is obtained by setting

$$c = \hbar = k_B = 1$$

leaving only one unit. This is usually selected to be GeV, and thus everything is measured in GeV. E. g., times and lengths are then both measured in GeV^{-1}. In certain circumstances, GeV are exchanged for fm, which an be easily converted with the rule

$$\hbar c = 1 \approx 0.1973 \text{ GeV fm}$$

and thus energies are then measured in inverse fm.

These units are the standard units of particle physics. No relevant publication, textbook or scientist uses any different units, except for occasionally using a different order, like MeV, being 10^{-3} GeV, or TeV, being 10^3 GeV. Also, other dimensionful quantities are then expressed in these units. E. g. Newton's constant is roughly 1.22×10^{19} GeV.

While astronomy has similarily also an adepted system of units, which is different from particle physics, this system is mainly driven by scales of the universe. As will be seen, any theory related to a common foundation will be more like particle physics, also for the regime of gravity, especially with Newton's constant above, will be expressed in GeV.

This is, of course, as unsuited to describing the solar system as it is to describe biology, but for teh purpose at hand it is very convenient.

[1] It would be even possible to abandon this scale, but it turns out to be helpful to most people to retain a single scale.

1.3 Particle physics

1.3.1 The purview of particle physics

Particle physics is the science of the smallest constituents of matter, and how they interact. In a sense, it evolved out of physical chemistry, where it was first resolved that all chemical substances, like molecules, are made out of set of chemical elements, of which we know currently roughly 120. At the turn of the 19th to the 20th century, it was then found that the atoms themselves were not elementary, but rather had constituents - a single nuclei, which was characteristic for the element, and a number of indistinguishable electrons. The latter number was in turn uniquely fixed by the nucleus.

In the early 20th century it was then found that the nuclei themselves are not elementary, but were made up out of just two types of constituents, the protons and neutrons, commonly denoted as nucleons. Again, and as will be described later in detail, these nucleons are not elementary, but are made up out of the quarks. At the current time, we do not know, whether either quarks or electrons do have a substructure, but substantial effort is invested to find out.

As is visible from this short historical remark, the study of elementary particles has changed subject many times over the course of time. Today, elementary particle physics is considered to be the study of the most elementary particles known to date, their interactions, and whether there could be even more elementary constituents. Today, such studies are inseparable linked to quantum mechanics, as quantum effects dominate the world of the elementary particles.

It is this field which will be presented in the first part of this lecture. After some more general remarks, the three known fundamental interactions, as well as the known particles will be presented. Their theoretical description together constitutes the standard model of particle physics. This theory is well established, and has been experimentally verified in the accessible range of parameters. But there are several reasons which prove that it can not be the final theory of particle physics. These will be briefly summarized later in the lecture. After adding gravity several candidates for extensions and combination of both will be introduced, and briefly discussed. Given their enormous numbers, and each of them warranting an own specialized lecture, these expositions can give only the roughest idea. It must furthermore be noted that fundamental physics is a very dynamical and living field, and the influx of new results, both experimentally and theoretically, is large. Hence, this material can only be taken as a snapshot of our current understanding.

As the standard model of particle physics in its current form has been theoretically established at the turn of the 60ties and 70ties of the 20th century, with the final major theoretical touches added in the early 80ies. However, the latest experimental results for the standard model included in this lecture are from the year 2021, and at the time of writing, a few input quantities still remain to be experimentally determined, as will be discussed below. Furthermore, the formulation of a theory is one problem. Determining all its consequences theoretically is a much harder problem, which is in many respects still unsolved today for the standard model. We are far from the point where we can compute any given quantity in the standard model to any given accuracy. Furthermore, experimental results are never exact, and have errors. So, in many cases we know for deviations from the standard model only upper limits, and cannot exclude something is different. Thus, our knowledge of the standard model is in continuous motion. The best repository of the current knowledge is from the particle data group, which is accessible for free at pdg.lbl.gov.

Note that particle physics without full-fledged quantum-field theory can necessarily often only give hints to the origin of many phenomena. The main aim is to make you acquainted with the general structure, with concepts, and with ideas. A full understanding of the details is necessarily relegated to quantum-field theory, and, in part, advanced quantum field theory courses. Wherever possible, I will try to give as much insights from these, without going into the technical details, skipping calculations for hand-waving arguments. As a result, with the knowledge of this course, you will be able to follow an experimental or theoretical overview talk at a mixed conference, but it will not be sufficient to follow a specialist talk in a parallel session, or likely an overview talk at a purely theoretical conference. Nonetheless, even in the latter cases, many of the ideas and underlying physics should sound then at least familiar to you. Eventually, the aim of this is to prepare you for a full quantum-field theory course on particle physics.

1.3.2 Elementary particles

From the point of view of elementary particle physics an elementary particle is any particle which, to the best of our knowledge, has no internal structure and is point-like. In quantum mechanics the Compton wave-length is often used to characterize the size of a quantum-mechanical object. Especially, this implies that an elementary particle is much smaller than its Compton wave-length. E. g., for an electron it is roughly of the order of its inverse rest mass, i. e. $(511 \text{ keV})^{-1} \approx 400$ fm. The experimental upper bounds for its size is 10^{-7} fm. Today, we know of 37 particles, which are elementary in this sense, and these will be introduced during this lecture.

On the other hand, there are composite particles, i. e. particles having an inner structure, which consists out of two or more of the elementary particles. Nucleons are an example for these. Of course, composite objects can consist out of even more composite objects, like nuclei, atoms, or molecules. Composite objects are also called bound states, as they are formed from bound together constituents. In the course of history, many instances are known that particles previously recognized to be elementary have actually been composite. Today, nobody would be too surprised if any of the elementary particles known would, in fact, be composite. This has happened too often.

1.3.3 Fermions and bosons

Since there exist elementary particles which are massless, like the photon, they move necessarily with the speed of light. Thus, any adequate description of particle physics requires the use of special relativity, and only under very special circumstances is it possible to obtain reasonably accurate approximate results by using only non-relativistic physics. Besides this necessity, the union of quantum mechanics and special relativity, called quantum field theory, implies many remarkable and very general results.

One such result is the fundamental distinction of particle types into bosons and fermions. Bosons obey Bose-Einstein statistics; this implies that there can be arbitrary many of them in any given state of fixed quantum numbers. On the contrary, fermions obey Fermi-Dirac statistics and have to adhere to the Pauli principle: In any given state with fixed quantum numbers can only be a single fermion.

The spin-statistics theorem, which is a very basic property of all experimentally verified field theories, and of most hypothesized ones, implies a deep relation between the statistical properties of particles, the Lorentz group, and the spin of a particle. It essentially states that for any Lorentz-symmetric, i. e. special relativistic, quantum theory in more than two (one space and one time) dimensions every particle obeying Bose-Einstein statistics has integer spin, and any particle obeying Fermi-Dirac statistics has half-integer spin. There are no further possibilities for the spin values. Thus, bosons have integer spin, and fermions have half-integer spin. The electron, e. g., is thus a fermion, with its spin of 1/2. The photon, with its spin of 1, is therefore a boson.

Note that spin is an intrinsic property, like rest mass, of an elementary particle. It is not something which can be constructed from any properties, and has to be determined in experiment, at least inside the standard model. For a bound state, however, the spin can be computed as a function of the spins of its constituents.

A further important consequence is that the wave-function of a system with n particles has to be antisymmetric under the exchange of two fermions, while it is symmetric under the exchange of two bosons. This already illustrates the Pauli-principle. If there would be two particles of the same quantum numbers (and thus of the same type, since type is a quantum number) at the the same place, their wave-function must change sign under an exchange. At the same time, they are identical, and thus numerically it cannot change. Since the only quantity which at the same time changes and does not change sign is zero, any such state has to be zero when the particles are at the same place, and thus they cannot be - this is the Pauli principle. On the contrary, for bosons the wave-function remains unchanged under an exchange of two identical bosons.

1.3.4 Particles and anti-particles

Another fundamental consequence of quantum field theory is the existence of anti-particles, first noted, and experimentally verified, in the 20ies and 30ies of the 20th century. The basic statement is that for any particle there exists an anti-particle of the same mass and spin, but otherwise opposite (e. g. electric) charges. For the electron, there is a fermion with also spin 1/2, positive electric charge, and the same mass: The positron. Only if a particle has no properties except for mass and spin (like the photon) there is no anti-particle, as the anti-particle would be again the particle itself.

As a further consequence, if a particle and its corresponding anti-particle meet, they can annihilate each other. If, e. g., an electron and a positron meet, they can annihilate each other, which will result in two or more photons. This does not need to happen instantaneously. Electron and positron can also first form a bound state, positronium, which has a finite life-time. It decays because of the annihilation of the two particles, decaying into two or three photons, depending on how the spin of the electron and anti-electron had been aligned with respect to each other.

The simplest way in which anti-particles appear can be gleaned from the mathematical description of non-interacting spin 0 particles of mass m. As any particle, they are actually described by a (complex) field $\phi(x)$ rather than a wave-function. They can be described by the Lagrange density

$$\mathcal{L} = (\partial_\mu\phi)^\dagger\partial^\mu\phi - \frac{m^2}{2}\phi^\dagger\phi.$$

The corresponding Euler-Lagrange equation is the Klein-Gordon equation

$$\partial_\mu\frac{\delta\mathcal{L}}{\delta\partial_\mu\phi^\dagger} - \frac{\delta\mathcal{L}}{\delta\phi^\dagger} = \partial^2\phi + m^2\phi = 0. \tag{1.3}$$

In momentum space, this equation is solved immediately by plane waves $\exp(ikx)$, with $kx = Et - \vec{p}\vec{x}$. Plugging them in yields for a particle at rest, $\vec{p} = 0$, $E^2 = m^2$, thus fulfilling the relation between the rest mass and the energy in the rest frame of special relativity. However, this result permits besides the usual solution $E = m$ also the solution $E = -m$. Such a negative energy solution makes no sense in ordinary quantum mechanics.

In quantum-field theory, where the Klein-Gordon equation is promoted to an equation for operators rather than for wave functions, this second solution can be associated with the anti-particle. Thus, even this simplest case already harbors particles and anti-particles. Note that the quantum number in which particle and anti-particle here differ is one which stems from the fact that the Klein-Gordon equation is invariant under the multiplication with a phase $\phi \to \exp(i\phi)\phi$, a so-called U(1) symmetry. Particles and anti-particles can differ in their quantum numbers for this replacement. Especially, an anti-particle is connected with complex conjugation, and its phase will rotate opposite to that of the particle.

Since this is a very fundamental prediction of our understanding of particle physics, great experimental effort is invested to check whether particles and anti-particles really have the same properties, e. g. the same mass. Particular attention has been devoted to produce anti-atoms, and measure their spectrum, as this is a very sensitive test of this theoretical prediction. So far, no deviation has been found. Still, although the production of anti-matter is tedious, and yields are measured in individual atoms, rather than grams, these tests remain the most sensitive ones of the foundations of particle physics.

1.3.5 Identical particles

One surprising fact was that elementary particles of a specific kind - no matter whether composite or point-like- are actually identical. They have exactly teh same mass, charges, etc.. Each electron has, within measurement uncertainties, exactly the same mass and electric charge.

This appears surprising. Why should there bo no variation. Quantum mechanics and classical mechanics do never require this to be the case. It is again the combination of quantum mechanics and special relativity, which explains this: Quantum field theory.

In quantum field theory there is no longer the concept of a particle. Rather, everything is described by fields, similar to the eltric and magentic ones in classical electrodynamics. What is perceived as a particle at longer distances is a(n exponentially) localized spike in the field. Because for every kind of particle there is only a single field, all particles of this kind are identical. Thus, the picture of quantum field theory is that there is only a handful of fields, which fill the whole universe, and all particles are jsut excitations of these fields. While this may seem strange at first, this is not that different from the idea of sending directed electromagnetic messages or laser pulses in classical electrodynamics. Just that it pertains also to objects where the particle nature of their wave-particle duality is more in the foreground in classical physics.

However, the picture of particles serves even in a quantum field theory often as a useful proxy, and often very good approximation. It will tehrefore be extensievly used throughout.

It is only by having suitable separations of these spikes that yields an apperance of actual particles. If particles come together what happens is that there is no longer distingusihable particles, but rather some broad excitation of the field. This is an interaction to be discussed next.

1.4 Interactions

The annihilation of electrons and positrons into photons is an example for an interaction, in this case an electromagnetic interaction. Writing on a black board or not falling through the floor are also examples of electromagnetic interactions. The earth orbiting the sun is an instance of a gravitational interaction. Generically, anything, except the Pauli principle, what makes a particle aware of the presence of another particle is classified as an interaction. This is called a dynamical effect. The propagation of particles, on the other hand, is just a kinematical effect.

Besides the knowledge of all the elementary particles, it is also necessary to know the interactions of them with each other to write down a theory describing them. Such an interaction is not necessarily restricted to connect only two particles. The maximum number known so far is a four-particle interaction, and any interactions involving more particles than four[2] can be broken down to those with less particles. This does not mean that it is not possible to have more; theoretically, this is possible, but there is no experimental evidence that it is needed so far. Also, there can be interactions which appear at a coarse scale to be an interaction involving more than four particles, but on a fine scale this is not the case.

There are two particular properties of interactions, which deserve a special mentioning. One is that interactions are not totally arbitrary. Rather they only occur between specific particles. E. g. the aforementioned annihilations only proceeds with one electron, one positron, and two or more photons. It will not occur with two electrons and two photons, or one electron, a proton, and two photons. The reason are conservation laws, in this case the one of electromagnetic charge. Throughout many more conservation laws will be encountered, and any interaction will strictly respect any exactly conserved quantity.

However, mass is not a conserved quantity. As a consequence of the famous Einstein equation $E = m$ (note the natural units!), energy and mass can be freely converted into

[2]Given the idea of identical particles as a field excitation, this seems blurry. However, what is meant can be given a very specific meaning, but then needs to abstract from the notion of particle.

each other. Thus in the positron-electron-annihilation process, the total rest mass of both particles, roughly 1 MeV, is converted into the kinetic energy of the photons, which are themselves massless. On the other hand, colliding two particles with a large amount of kinetic energy will permit to create new particles, which have at most a total mass equal to the energy (and rest mass) of the original particles. This is of central importance to scattering experiments to be discussed in chapter 2.

Before moving on to them, there is one more noteworthy kind of interaction. Quantum mechanics forbids to measure energy and time at the same time arbitrarily precisely[3], $\Delta E \Delta t \geq 1$. Thus, it is possible for a brief amount of time to have more energy in the system. For this duration, it is possible to create also particles with this additional energy, and thus more rest mass than originally available as energy, though they have to be destroyed at the end of this period. Since these live only for this short amount of time, they are not real, but are called virtual particles. Such particles are often denoted by an asterisk, *.

Such virtual pairs can either appear out of the vacuum, or can be emitted and reabsorbed by other particles. In both cases there have been many experimental confirmations of these processes, e. g. the Casimir effect for the first and the Lamb shift for the latter. Such virtual particle fluctuations are today used in precision experiments to access particles with masses larger than those which can be directly created using available energies. This will also be discussed later.

1.5 Gravity and quantum gravity

The other fundamental theory known today besides particle physics is the classical theory of general relativity. It describes a single interaction, gravity, and matter only in the form of its energy-momentum tensor. It has thus much less degrees of freedom than current particle physics. Nonetheless, it is vital in the description of macroscopic objects, from our solar system to the universe as a whole. It is also a field theory, i. e. gravity is described by a field, similar to electrodynamics.

While it is technically much more involved than classical electrodynamics, it follows essentially the same structure. The Einstein equations are the equiavlent to Maxwell's equation, and are partial differential equation. The energy-momentum tensor takes the place of the electric four-current, and like the homogenous Maxwell equations there are geometrical constraints. However, it is technically much more demanding, and describes

[3]Actually, this is not an elementary relation like the Heisenberg uncertainty principle, but follows from identifying particles with wave packets, as well as from the uncertainty principle proper.

phenomena much more removed from everyday experience. This makes it somewhat harder to access. Details will be given in chapter 7.

However, because matter is quantized, so is the energy-momentum tensor. It is thus necessary that ultimately also gravity becomes a quantum theory[4], quantum gravity. This would unify general relativity with the standard model of particle physics. However, an estimate of at which scales quantum gravity effects become observable yields about 10^{19} GeVin energy or 10^{-35} m, far removed from anything experimentally accessible in teh foreseeable future. Thus, only consistency arguments can be used right now to determine the theory of quantum gravity. Though a few other ideas also exist, see chapter 9.5. As a consequence of this, and of the technical complexity already of general relativity, there is not yet agreement on how quantum gravity could look like, much less experimental support for any option.

1.6 Bottom-up, top-down, and effective theories

There have been two facts about any fundamental theory of nature over the course of history.

One was that they predicted their own failure, usually due to the apperance of divergencies. This range from the gravitational energy of point particle and hohlraumstrahlung and ends at the short-distance behavior of quantum field theories, to be discussed in more detail starting from section 3.7, and the apperance of singularities at the center of black holes in general relativity. These divergncies always signalled the breakdown of a theory starting from a certain point on, and the need to have a different theory.

The other was that the new theory always[5] contained the previous one as a limiting case. Mechanis was found as a small-speed limit of special relativity, which in turn is a small-curvature limit of general relativity. Mechanics is also a long-distance limit of quantum mechanics, and the unification of mechanics, electrodynamics, and special relativity a long-distance limit of quantum field theory.

Likewise, specialized fields can be derived from such theory. Thermodynamics is the limit of many particles of mechanics, optics the long-wavelength limit of electrodynamics, continuum mechanics the limit of small distances between particles in mechanics, and many other.

[4]This appears to be the easiest solution, where 'easy' is still very, very hard.

[5]At times there have been more than one proposal for theories, which were in accordance with experiments at that time. By now, it has been recognized that with more expeirments one would have been singled out.

As a consequence, physics is now seen as a hierarchy of such theories, called effective theories. In addition, in their realm of applicability these effective theories are usually much simpler to apply then the underlying, more encompassing theories. In fact, more complicated phenomena can often only be described with adequate amount of work in the effective theories, even if it is in principle possible in the most fudnamental ones.

At the moment, the union of general relativity and particle physics as a quantum gravity standard model forms, to the extent as it is yet known, the most fundamental theory, of which all others can be derived. Our best approximation to nature. However, due to the divergencies encountered, it is pretty likely not the last one, and there is at least one, but likely many more, hierarchy levels to be found. It is an interesting question, also from the point of view of the philosophy of physics, whether there can be an end to this hierarchy: A unified theory which explains everything, including itself. It is not clear, whether such a thing is even conceptually possible, and it remains one of the most fundamental challenges in science.

When searching for a new, more encompassing, theory, there are two different pathways. One is the so-called bottom-up approach. In this case, slight (or, in lucky cases, not so slight) experimental deviations from the predictions of existing theories are used to infer the structure of the underlying theory. In the top-down approach, based on a guess at the possible underlying principle an encompassing theory is constructed. Then predictions for deviations are made, and then expeirments will decide the fate of the proposal.

During the 20th century an abundance of experimental deviations made the bottom-up approach extremly successful. Especially, if a deviation[6] is present, this guarantees a discovery. In the later 20th century up to now deviations became scarce. This led to a stronger emphasis of top-down approaches. Though there still exist observed deviations, see chapter 8, this has not yet led to a success. It is thus an interesting situation right now in the search for the fundamentals of physics, not so dissimilar from the end of teh 19th century: It is known that something is amiss, but it is not clear how.

[6]Provided, of course, the theoretical calculations which are used, are accurate enough. Which, with a finite amount of ressources, is not awlays possible.

Chapter 2

Scattering experiments

The primary tool to understand (particle)2 physics today experimentally are scattering experiments, i. e. letting particles collide with each other. The reason has been alluded to before: When two particles collide they interact, and their kinetic energy can be used to create new particles. The higher the energy, the more massive particles can be created. Furthermore, because of the DeBroglie relation that wave-length and momentum are related by $\lambda = 1/p$, only high energies permit to resolve very small structures, and therefore permit to investigate whether a given particle has a substructure.

Of course, due to the presence of virtual particles, it would in principle be possible to also investigate everything using low-energy collisions or even measurements of static properties of a particle. However, as will be seen later, the required precision to resolve new particles in most cases drops with increasing energy. Therefore, to access new particles in low-energy collisions requires very precise experiments. For some cases, particular effects reverse the situation. As a consequence, today both low-energy precision experiments and high-energy collisions work hand in hand together, both having their own importance.

The formulation of scattering experiments is thus central to particle physics. Almost all discoveries in particle physics have been made with such experiments. As a consequence also most modern theoretical tools have been developed and optimized to describe scattering experiments. Thus, at least an elementary understanding of scattering processes is a necessity in particle physics.

2.1 Transitions

The basic concept behind scattering experiments is that it is possible to perform a transition from one (initial) state $|\alpha\rangle$ to another (final) state $|\beta\rangle$, characterized by a transition matrix element $\langle\beta|\alpha\rangle$. Scattering itself is just a tool with which this transition is conveniently realized. It is thus worthwhile to briefly repeat the basic ideas behind transition and scattering in quantum mechanics.

Usually, particle physics theory can be separated in a free part and an interaction part. The former can be solved exactly, while the full theory is usually accessible using approximations. One of the most successful approximations is perturbation theory. Since almost all non-perturbative approaches are generalizations of perturbative approaches, it is always a good starting point to first understand the perturbative one, and then move on to the more complicated non-perturbative treatment. Hence, assume that α and β differ only slightly, in a sense to be made more precise below.

Hence, assume that the Hamilton operator can be split into the free part H_0 and the interaction part H_{int}, with H_{int} being only a small perturbation. It is then possible to expand any state into eigenstates[1] of H_0 as

$$|i\rangle = \sum_n a_n^i \exp(-iE_n t)|n\rangle, \tag{2.1}$$

where the a_n^i are the nth (time-dependent) expansion coefficient of state i for the eigenstate $|n\rangle$ of energy E_n. A transition matrix element between two states can then be determined, using the orthogonality of the eigenstates.

To obtain the influence of the interaction on a state $|i\rangle$, insert the expansion (2.1) into the Schrödinger equation

$$i\sum_n \dot{a}_n^i \exp(-iE_n t)|n\rangle + \sum_n E_n a_n^i \exp(-iE_n t)|n\rangle = \sum_n a_n^i \exp(-iE_n t)(H_0 + H_{\text{int}})|n\rangle,$$

which shows that this introduces a(n additional) time-dependence in the expansion coefficients. To isolate a single coefficient, project with a second eigenstate $\langle N|$, yielding

$$\dot{a}_N^i = -i\sum_n \langle N|H_{\text{int}}|n\rangle a_n^i \exp(i(E_n - E_N)t),$$

which shows that the coefficients have a complicated oscillatory pattern.

For a perturbative setting, assume that all the coefficients are small, expect for the one α, which thus is almost constant in time. By a rotation of base it is always possible to achieve this that α is one of the eigenvectors. Then the sums collapse to leading order, and an ordinary differential equation ensues, which can be solved directly, to yield

$$a_N^\alpha = \delta_{N\alpha} - i\sum_n \langle N|H_{\text{int}}|\alpha\rangle \left(\frac{1 - \exp(i(E_N - E_\alpha)t}{i(E_N - E_\alpha)}\right).$$

[1] In a fully relativistic theory subtleties arise, which are usually practically irrelevant.

This shows how slowly additional eigenstates become populated, while the original one becomes depleted. Of course, this approximation will in general break down, when the coefficient of α becomes of similar size than the other ones.

The probability for a transition from α to N is then

$$P_{N\alpha} = 4|\langle N|H_{\text{int}}|\alpha\rangle|^2 \frac{\sin^2 \frac{(E_N - E_\alpha)t}{2}}{(E_N - E_\alpha)^2}.$$

Hence, at any fixed time, transitions to states with vastly different energies are strongly suppressed, which is consistent with the assumption of a small perturbation. In fact, for times T much shorter than half a period, essentially only states with energy difference $\Delta E \approx 2\pi/T$ will be populated. Hence, it can be assumed that the matrix element is essentially independent[2] of N, and it can be replaced by $|\langle\beta|H_{\text{int}}|\alpha\rangle|^2$.

Summing over N will provide the total transition probability. Since usual particle physics experiments will be on unbound particles, which are scattered, the energy levels can be assumed to be a continuum. Hence, the sum turns into an integral, weighted by the density of states, yielding as total transition probability

$$P(T) = 2\pi T|\langle\beta|H_{\text{int}}|\alpha\rangle|^2 \left.\frac{dN}{dE}\right|_{E=E_\alpha\approx E_\beta} \tag{2.2}$$

where the last factor is the density of states at the energy, also known as phase space $\rho(E)$. This formula is also known as Fermi's golden rule.

Under the assumption of weak interactions, the phase space is just the one of free, non-interacting particles, which is given for n particles in the state β, the final state, by

$$\rho_n = \lim_{V\to\infty} \frac{V^{n-1}}{2\pi^{3(n-1)}} \frac{d}{dE} \int d^3p_1 ... \int d^3p_{n-1}\delta_E,$$

where V is an auxiliary volume in which to quantize the energy levels, and the integrals are over the spatial momenta of the n particles, which are constrained by a suitable function δ_E to satisfy energy conservation, such that the total energy is E.

The structure of a scattering probability is always like the expression (2.2), even in much more complicated cases: Some numbers, the transition matrix element, which may be more complicated beyond leading order, and a phase space contribution of the final states. If part of the final state is not observed, it has to be summed or integrated over. E. g. if the particles in the final state have spin, which is not observed, the result has to be averaged over all possible spin orientations. Concerning the initial state, in most of this text it will be assumed that is known as precisely as quantum mechanics permits.

[2]Otherwise a perturbative treatment is not justified.

2.2 Non-relativistic

Specializing now from general transitions to scattering experiments, it is useful that many of the necessary concepts can already be understood non-relativistically. This offers a simpler formalism.

2.2.1 Elastic scattering

The simplest case is that of two incoming particles colliding with each other, and leaving as the outgoing particles unchanged. This case is known as elastic scattering. A very suitable coordinate system to describe this is the center-of-mass coordinate system, in which the total momentum of the two incoming particles is zero, $\vec{p}_1 = -\vec{p}_2$ with equal length p. Other coordinate systems, like the laboratory frame, with non-zero center-of-mass momentum can be obtained using a Galileo transformation (or later a Lorentz transformation), and therefore will not be discussed here further. Furthermore, if particles have a spin, things are also more complicated, but this will be avoided for now. Especially, in the following always bosons will be assumed, and the Pauli principle will not play a role.

The initial state is thus totally described by the type of particles and their properties, say two particles with masses m_1 and m_2, and their two momenta. Non-relativistically, energy and momentum conservation hold separately, and the initial state is completely characterized by the quantities

$$\begin{aligned} E &= \frac{p_1^2}{2m_1} + \frac{p_2^2}{2m_2} \\ 0 &= \vec{p}_1 + \vec{p}_2. \end{aligned}$$

Important here is the absence of any potential energy. In scattering experiments, it is assumed than in the initial and final state the particles are so far separated that they do not influence each other. This is called asymptotic states. For electromagnetic and gravitational interactions, which decay as $1/r$, this approximation is certainly reasonable. It is also found to hold true for all particles and interactions, which can so far be prepared as initial states, or detected as final sates.

The final, or outgoing, state, has to have the same total energy, and the same masses. Thus, the total final momenta $\vec{q}_i$ has the same size, $q_i = p_i$, but may have a different direction. However, there is only one undetermined quantity, since the scattering is coplanar: There is a possible scattering angle, which can be defined between any incoming and outgoing momenta, e. g. $p_1 q_1 \cos\theta = \vec{p}_1 \vec{q}_1$. There are many conventions in use, between which two momenta the scattering angle is measured, and in which direction.

This scattering angle θ depends on the interaction with which the two particles scatter. Measuring it and comparing it to the prediction is one of the possibilities on how a theory can be falsified.

2.2.2 Cross section

In a quantum theory, a scattering process will be statistically distributed, and therefore the measurement of an individual scattering angle is meaningless. Thus, experiments are repeated many (many billion) times to obtain a distribution of scattering angles. The number of particles scattered into a solid angle $d\Omega = \sin\theta d\phi d\theta$ is then defining this cross section as

$$\frac{d\sigma}{d\Omega} = \frac{dN}{d\Omega}.$$

In practice, there are various quantities, which can be derived from the cross section, which are of particular importance. One is the total cross section

$$\sigma_t = \int \frac{d\sigma}{d\Omega} d\Omega,$$

which may still depend on further parameters, like the total center-of-mass energy. The others are differential cross sections

$$\sigma_x = \frac{d^2\sigma(\theta, \phi, x)}{d\Omega dx},$$

where x can be any kind of variable, including also θ and ϕ themselves, on which the cross-section can depend. Also higher order differentials could be useful, so-called multiple-differential cross sections. Of course, differential cross-sections can also be defined from the total cross-section.

The unit of a cross section is an area. A useful unit to measure it in particle physics is barn, abbreviated b, corresponding to 100 fm^2. Nowadays, usually interesting cross-sections are of the order of nb to fb.

Predicting such cross sections is one of the most important tasks when one wants to test a given theory against modern particle physics experiments. However, this is quite an indirect process. The cross-section is calculated as a function of the parameters of a theory, which includes also the type and number of elementary particles. This result is then compared to the experimental result, an by comparison of how the cross-section is then depending on external parameters (e. g. the angle or the type of particles involved), the theory is either supported or falsified.

This becomes complicated in practice, if the cross section has contribution not only from the theory in question, but especially also from other known origins. E. g., when searching for yet unobserved physics, one has to subtract any contribution of known physics from the measured cross-section. Especially if the known contribution is the dominant one, this is difficult, as it is practically impossible to calculate it to arbitrary precision. It is also not possible to derive it from experiment; there is no possibility to continue unambiguously an experimental result from one case to another, without performing a theoretical calculation. Thus, this background reduction of known processes has become one of the major challenges in contemporary particle physics.

2.2.3 Luminosity

The definition of a cross-section yields immediately another figure of merit for a modern experiment, the luminosity. Technically, it is defined as the number of particles in two beams, N_1 and N_2, which interact in an (effective) area A with frequency f,

$$L = \frac{N_1 N_2 f}{A},$$

i. e. this is the number of processes per unit time and unit area. In modern particle physics experiments, the beam is usually bunched into n packages, which act like there would be several beams, and thus multiply the luminosity by n.

More important is the integrated luminosity,

$$\mathcal{L} = \int^{T} dt L,$$

that is how many collisions per unit area occur during an experiment of temporal length T. From this quantity, the expected number of events N from a process with a cross-section σ can be calculated as

$$N = \sigma \mathcal{L}.$$

Especially, to obtain one event occurring with cross-section σ during the time of the experiment, an integrated luminosity of $\mathcal{L} = 1/\sigma$ is necessary. This permits to directly estimate whether an experiment is able to measure a process at all. Typical values for integrated luminosity at the LHC are 25 fb^{-1} until the end of 2012, and about2 300 fb^{-1} until 2020.

A concept which will become important later on is the so-called partial (or later parton) luminosities. Assume that the beams consist not out of a single type of projectiles, but several. Then one can define for each possible pairing an individual partial luminosity.

2.3 Relativistic

2.3.1 Repetition: Relativistic notation

A non-relativistic description is, unfortunately, not sufficient in particle physics. There are three reasons for this limitation. One is that there are massless particles, which necessarily move with the speed of light, and therefore require a relativistic description. The second is that many regularities of particle physics are very obscure when not viewed from a relativistic perspective. Finally, in the reactions of elementary particles the creation and annihilation of particles, a conversion of energy to mass and back, is ubiquitous. Only (special) relativity permits such a conversion, and is therefore mandatory.

To just repeat the most pertinent and relevant relations, the starting point are the Lorentz transformations for a movement in the 3-direction,

$$\begin{aligned} x_1' &= x_1 \\ x_2' &= x_2 \\ x_3' &= \gamma(x_3 - \beta x_0) \\ x_0' &= \gamma(x_0 - \beta x_3) \\ \gamma &= (1-\beta^2)^{-\frac{1}{2}} \end{aligned}$$

which leave the length $x^2 = x_0^2 - \vec{x}^2$ invariant. The quantity x denotes the corresponding four-vector with components x_μ in contrast to the spatial position vector $\vec{x}$ with components x_i. Note that $\beta = |\vec{\beta}|$ is in natural units just the speed.

The scalar product of two four-vectors can be obtained using the metric $g = \mathrm{diag}(1, -1, -1, -1)$, i. e.,

$$x^2 = x_\mu x^\mu = x_\mu g^{\mu\nu} x_\nu,$$

i. e., the metric lowers and raises indices, which distinguish covariant and contravariant vectors. Though there are profound geometric differences between both, these have essentially no bearing on particle physics in the context of the standard model of particle physics.

From this follows the definition of the relativistic momentum and energy,

$$\begin{aligned} p &= \gamma\beta m \\ E^2 &= \vec{p}^2 + m^2 \end{aligned} \tag{2.3}$$

where m is the (rest) mass of the particle, an intrinsic and immutable (at least within known particle physics theories) property of any given type of particle. The concept of a relativistic mass, defined as γm is actually not necessary nor relevant in particle physics, since rest mass, and spatial momentum completely characterize the state of a particle. Note that for $\vec{p} = \vec{0}$ the relation (2.3) implies $E = m$, and thus rest mass and energy in the rest frame are exchangeable concepts. Especially, this relation implies that mass and energy can be freely converted into each other in special relativity.

An important special case are massless particles, which always move at the speed of light, i. e. $\beta = 1$. In that case, $E = |\vec{p}|$ precisely. Note that γ is neither well-defined, nor needed, when describing massless particles.

The relation (2.3) is only true for real particles. The virtual particles stemming from quantum fluctuations introduced earlier do not fulfill it. Such particles are called off-shell, since the relation (2.3) defines a surface in momentum space, called the mass-shell. Particles fulfilling (2.3) are consequently called on-shell.

One further consequence relevant to particle physics is time dilatation, i. e. that the time (difference) τ' observed in a frame moving relative to the rest frame of the particle is longer than the eigentime τ in the rest frame by

$$\tau' = \gamma\tau. \tag{2.4}$$

Of course, this also implies that a moving particle experiences in its own rest frame a shorter time than outside. This will be very relevant for objects with a finite life-time: They exist the longer the faster they are. This has especially consequences for the experimental accessibility of massive particles.

Elastic scattering proceeds then in the relativistic case just as in the non-relativistic case. The only change is that the separate conservation of spatial momentum and energy is replaced by the conservation of four momentum,

$$p_1 + p_2 = q_1 + q_2.$$

Still, the identities of the particles are conserved in an elastic scattering, and therefore their rest mass is not changed. Also, there is still only one scattering angle characterizing this (also still coplanar) reaction.

However, scattering angles are cumbersome quantities, as their Lorentz transformations are rather complicated. As a consequence, more suitable quantities are used in practice. These are based on the concept of the rapidity ϕ. The rapidity of a particle is defined as[3]

$$\phi = \tanh^{-1}\frac{|\vec{p}|}{E} = \frac{1}{2}\ln\frac{E + |\vec{p}|}{E - |\vec{p}|},$$

and thus relates to the energy and momentum as

$$\begin{aligned} E &= m\cosh\phi \\ |\vec{p}| &= m\sinh\phi, \end{aligned}$$

[3]Note that the name rapidity sometimes also refers to the boost vector in a Lorentz transformation.

and hence is infinite for massless particles. It is related to the contraction factor γ as $\gamma = \cosh\phi$, i. e. the rapidity is added when increasing the speed of a frame.

In particle physics experiments, where particles are collided head-on, there is a distinct axis, usually chosen to be the z-axis. In this case, a rapidity can be defined with respect to the beam axis by replacing the momentum p with the momentum p_L along the z/beam-axis. Though this conserves only some of the useful properties under Lorentz transformation of the true rapidity, these are the most useful in practice, and often this modified rapidity is therefore used instead. In contrast, the momentum perpendicular to the beam axis is usually called transverse momentum p_T, sometimes also $p_\perp$ and called p-perp.

Defining now the scattering angle θ with respect to the beam axis, it is finally possible to introduce the pseudorapidity η,

$$
\begin{aligned}
\eta &= -\ln\tan\left(\frac{\theta}{2}\right) \\
\theta &= 2\arctan e^{-\eta}.
\end{aligned}
$$

Using the outgoing momentum of the scattered particle, the pseudorapidity is given by

$$
\eta = \frac{1}{2}\ln\left(\frac{|\vec{p}| + p_L}{|\vec{p}| - p_L}\right).
$$

If the energies and momenta are large compared to the rest mass, this essentially coincides with the modified rapidity, and is therefore quite often used interchangeably.

However, an interesting new possibility are now inelastic scatterings where particles are produced or destroyed[4].

2.3.2 Inelastic scattering

In particle physics, it is no longer necessary that the particles scattered retain their identity after the scattering process. Especially, they can be transformed into a number of other particles, a so-called inelastic collision. Of course, the final products must still conserve 4-momentum. Thus,

$$
p_1 + p_2 = \sum_n q_i.
$$

[4]In principle, e. g. by breaking up into more than two fragments, there are also non-relativistic inelastic scattering processes possible, where the total mass is conserved. Though this plays an important role in, e. g., chemistry, this possibility will be ignored here.

This implies especially

$$\begin{aligned}\sqrt{\vec{p}_1^2+m_1^2}+\sqrt{\vec{p}_2^2+m_2^2} &= \sum_i \sqrt{\vec{q}_i^2+m_i^2}\\ \vec{p}_1+\vec{p}_2 &= \sum_i \vec{q}_i.\end{aligned}$$

In the center-of-mass frame, where the initial 3-momentum vanishes, it is thus in principle possible to transform the rest energy and the kinetic energy of the initial particles only into the rest energy of a number of new particles. This is called an on-resonance production. Especially, it is possible to create particles which are heavier than the original ones, provided the kinetic energy is large enough. This explains why higher and higher energies are required in particle experiments to obtain heavier and heavier particles.

Of course, this can be generalized to more than two incoming particles. However, in practice it is almost impossible in an experiment to coordinate three particles such that they collide simultaneously with any reasonable rate. Hence, this will not play a role in the following. It is not irrelevant, though. If the number of collisions becomes much larger than currently technically achievable, this can happen for an appreciable fraction of the events. Inside the sun such huge numbers are possible, and then three-body interactions become important for the thermonuclear processes. In fact, without a certain reaction of three helium nuclei to a carbon nucleus, the production of heavier elements in the sun would not occur in the way it does in nature.

This also yields another important way of how to deconstruct cross sections. Assume a collision of two particles of type x as the so-called initial state. As a quantum mechanical process, the reaction will not only lead to a single outcome, a single final state, i. e. only one set of final particles. If it is possible, given the conservation of four-momentum and any other constraints, that there are more than one possible outcomes y_i, they all will occur. Their relative frequency is given by the details of both the reaction and the underlying interactions. These rates n_i for n final states yield the partial cross-sections,

$$\sigma_{2x\to i}=\lim_{n_i\to\infty}\frac{n_i}{\sum_i^n n_i}\sigma.$$

Partial cross sections can also be defined in a broader sense. Instead of having precisely defined particles in the final state it is e. g., possible to define a partial cross-section for n particles of any type.

Partial cross-sections play an important role in experiment and theory alike. Measuring a total cross-section is called an inclusive measurement, while identifying some or all final particles and their properties are called exclusive measurements. At the same time, theoretical calculations permit to make a judicious choice of a final state in which the background due to known physics is small or vanishing, thereby increasing the signal-to-noise ratio. Determining partial cross-sections is therefore a very important task.

2.3.3 Partial wave decomposition

Of quite substantial importance is the possibility to decompose scattering events into partial waves associated with different values of angular momentum. This is on the one hand important to identify the properties of involved particles, but gives also interesting insights into bounds of various scattering process. The latter are, e. g., interesting as they were used to indirectly predict the presence of new particles and forces.

To develop the relevant concepts, start with the expansion of a plane wave ψ with momentum k in z-direction in incoming and outgoing spherical waves. This is an expansion in angular momentum eigenfunctions, i. e. Legendre polynomials P_l of angular momentum l,

$$\psi = e^{ikz} = \frac{i}{2kr}\sum_l (2l+1)\left((-1)^l e^{-ikr} - e^{ikr}\right) P_l(\cos\theta)$$

The first term is from the incoming waves, while the second from the outgoing.

The scattering process can alter only the outgoing waves, by modifying both its phase and its normalization. However, the later is bounded, due to the normalization. The resulting total scattered wave is hence

$$\psi_s = \frac{i}{2kr}\sum_l (2l+1)\left((-1)^l e^{-ikr} - \eta_l e^{2i\delta_l} e^{ikr}\right) P_l(\cos\theta), \tag{2.5}$$

where $0 < \eta_l \leq 1$ and δ_l is the so-called phase shift. The latter quantity is a very important quantity, as it is possible to determine from it the appearance of intermediate unstable particles, as will be discussed later.

Now, in the rest frame of the target particle, and if the incoming particle can be modeled by a plane wave, the difference between ψ and ψ_s is hence the wave-function of the outgoing particle. This wave-function can be used to define the scattering amplitude G

$$\psi_o = \psi_s - \psi = \frac{e^{ikr}}{r}G(\theta) = \frac{e^{ikr}}{r}\frac{1}{k}\sum_l (2l+1)\frac{\eta_l e^{2i\delta_l} - 1}{2i}P_l(\cos\theta),$$

which is related to the transition matrix element appearing also in (2.2). To obtain the result in the center-of-mass frame will require to perform a corresponding transformation, which will not be done here.

More interesting is that the scattering amplitude is directly linked to the differential cross-section. Since the norm of the wave-functions describe a probability for a particle to scatter into a certain solid angle $r^2 d\Omega$, the equality

$$\psi_o^* \psi_o r^2 d\Omega = |G(\theta)|^2 d\Omega = d\sigma$$

or

$$\frac{d\sigma}{d\Omega} = |G(\theta)|^2.$$

Hence, knowing the scattering amplitude is sufficient to describe the observed cross-sections.

Integrating over the solid angle delivers the full elastic cross-section

$$\sigma_e = \frac{16\pi^3}{k^2} \sum_l (2l+1) \left| \frac{\eta_l e^{2i\delta_l} - 1}{2i} \right|,$$

where in a fully elastic scattering $\eta_l = 1$. The absence of scattering will be at $\delta_l = 0$ and $\eta_l = 1$. Subtracting in a similar way the total incoming probability from the elastic outgoing probability, $\psi_o - \psi$, delivers the inelastic cross-section

$$\sigma_i = \frac{4\pi^3}{k^2} \sum_l (2l+1)(1 - \eta_l^2),$$

and both together the total cross-section

$$\sigma_T = \sigma_e + \sigma_i = \frac{8\pi^3}{k^2} \sum_l (2l+1)(1 - \eta_l \cos 2\delta_l) = \frac{4\pi}{k} \Im G(0). \tag{2.6}$$

The last equality follows from comparison, and is known as the optical theorem.

Since η_l is bounded, the cross-section is bounded, and this follows entirely from the conservation of probability, or, more general, unitarity. This leads to so-called unitarity bounds for the partial-wave cross-sections,

$$\begin{aligned} \sigma_e^l &\leq \frac{8\pi^3}{k}(2l+1) \\ \sigma_i^l &\leq \frac{4\pi^3}{k^2}(2l+1). \end{aligned}$$

These bounds have been used to infer the existence of a Higgs particle, which will be discussed in much detail later, with a mass below roughly 1 TeV prior to its discovery[5]

Using these bounds, as well as general theoretical considerations, it is possible to show that for all theories, which are experimentally relevant at the current time, the

[5]Strictly speaking, it was only possible to interfere the existence of something below 1 TeV, of unknown nature. Circumstantial evidence, however, pointed strongly to the Higgs being this something.

center-of-mass total cross-section cannot grow faster than $\ln^2\sqrt{s}$, where s is the squared center-of-mass energy. This is the so-called Froissart bound.

Finally, note that the quantity

$$f(l) = \frac{\eta_l e^{2i\delta_l} - 1}{2i}$$

can be used to characterize the scattering process. It describes a disc in the complex plane, the so-called Argand circle, of radius $\eta_l/2$. If the scattering is purely elastic, $f(l)$ will be on the boundary of the Argand circle, while it will be inside if the scattering is inelastic, i. e. $\eta_l < 1$.

The best studied examples for such cross-sections are so-called Drell-Yan processes. In such processes, a particle and an anti-particle collide, and are converted into a particle and an anti-particle of a different kind.

2.4 Formfactors

In practice, a very important way to describe cross-sections is by means of so-called form factors. Define q as the amount of momentum exchanged between two spin-less particles colliding into two other particles. Then a form-factor F is defined as

$$\frac{d\sigma}{d\Omega} = \frac{d\sigma}{d\Omega}_{\text{simple}} |F(q^2)|^2$$

The differential cross-section called 'simple' is the cross-section of two particles with point-like structure and exactly solvable interactions. Using Lorentz symmetry, it is possible to show that any influence of the internal structure can only depend on the momentum transfer q^2 between both particles, and therefore the deviation from such a structure can be parameterized by this single additional function, the form-factor. The interesting physics is hence entirely encoded in this form-factor.

Of course, if there are different particles involved, spin, or more than two particles in the final state, the structure is not as simple. However, Lorentz symmetry always permits to write the differential cross-section as a sum of products of trivial kinematic cross-sections and form factors. Hence, even in the general case, it is possible to parametrize a cross-section by a set of form factors. Since in this way all 'trivial' effects are removed, this is very useful to identify new effects.

2.5 Decays

In a sense a very particular type of inelastic processes is the decay of a single particle, i. e. a process with a single initial particle, and two or more final particles. A decay occurs if a massive particle can, in principle, be converted into two or more lighter particles. For this to be possible, the rest mass of the original, or parent, particle must be larger than the sum of the rest masses of the daughter particles, plus possibly a certain amount of energy to satisfy spatial momentum conservation. Furthermore, there may be additional constraints like that electric charge is conserved.

Since the decay of a particle is a quantum-mechanical process, it does not occur after a fixed decay time τ, but occurs at a time τ with an exponentially decaying probability, $\exp(-\Gamma\tau/2)$, where Γ is the so-called decay width. The value of the decay width is determined by details of the interaction, but is in most cases the larger the larger the difference between the rest mass of the initial particle and the total rest mass of the final particles is. Furthermore, the ratio Γ/m is used to characterize how unstable a particle is. If this ratio is small, the particle is rather stable, as it can traverse many of its Compton wave-lengths before decaying. Such particles are also known as resonances. If the ratio is close to or larger than one, the particle is not really behaving as a real particle, and there is no final consensus on whether such objects should be regarded as particles or not. Most particles well accessible in experiments have small widths.

However, quantum mechanically all possible outcomes can always be realized, just with differing probabilities. Therefore, if it is kinematically, and with respect to all other constraints, possible for a particle to decay into different final states, this will occur. Each possible outcome is called a decay channel, and is characterized by a corresponding partial decay width Γ_i, The sum of these widths together can be shown to yield the total decay width, which determines the life-time of the particle.

For the experimental measurement of the properties of resonances time dilatation (2.4) plays an important role. Since in experimental particle physics decay products are often produced at considerable speeds, their life-time in the experiment's rest frame is usually much larger than the life-time in their respective rest-frame. As a consequence, they can travel over considerable distances away from where the original particle has been produced, which permits to disentangle the processes of resonance production and decay. This also permits to much better determine the decay products, which in turn permit to identifying the particle which has decayed.

A simple, and heuristically, mathematical description of decays in quantum physics can be achieved in the following way. Though sketchy, the results agree with those obtained in a more rigorous treatment.

For a stationary state, the time-dependence of a quantum-mechanical state is given by the oscillatory factor $\exp(iE_0t)$ in the wave function $\Psi(x)\psi(t)$. This implies that the probability to find a particle is constant in time, as $|\psi|^2$ does not depend on the time anymore. If a particle should decay, this must change. This can be introduced by giving the energy a small imaginary component, $E_0 \to E_0 - i\Gamma/2$. Then the wave-function is damped in time, characterized again by the decay width Γ.

A decaying probability is however hard to interpret. It is better to perform a Fourier transformation in time. The Fourier-transformed wave-function becomes, dropping the spatial part $\Psi(x)$ which only acts as a spectator,

$$\psi(E) = \frac{1}{(2\pi)^{\frac{1}{2}}} \frac{i}{E - E_0 + \frac{i\Gamma}{2}}.$$

From this the (normalized) probability to find the particle with energy E can be obtained

$$P(E) = \frac{\Gamma}{2\pi} \frac{1}{(E - E_0)^2 + \frac{\Gamma^2}{4}}.$$

This shows that the additional imaginary part of the energy broadens the distribution in energy space of the particle. A resonance has thus no associated fixed energy, and has a so-called natural linewidth.

Such decays do not only occur for free particles. There are also known, e. g., in the context of atoms. If an electron in an atom is in an excited state, the atom will also decay into its ground state by emission of a photon, and the electron dropping back to the lowest possible level. In this case, E_0 will play the role of the energy of the excited atom, with $1/\Gamma = t_{1/2} \ln 2$ its life-time, or half-life $t_{1/2}$.

2.6 Cross-sections and particles

The formation of resonances yields, at least to leading order, a unique signature in the cross-section. Consider the situation that two particles of momenta p_1 and p_2 are collided, and react, elastically or inelastically, with each other, yielding two final particles with final momenta q_1 and q_2 with $p_1 + p_2 = q_1 + q_2$. If this process occurs such that the two original particles annihilate and form a resonance, which consecutively decays into the final products, it can be shown that the cross section behaves as

$$\sigma \sim \frac{1}{((p_1 + p_2)^2 - m^2)^2} \tag{2.7}$$

If the initial momenta are tuned such that the center of mass energy s fulfills $s = (p_1 + p_2)^2 = m^2$, the cross-section diverge, and thus the name resonance for the intermediate particles. Away from this condition, the cross-section quickly drops. Thus the formation of an intermediate resonance is signaled by a sharp peak in the cross-section. Since the relevant variable is the center-of-mass energy s, this is also called an s-channel process.

In contrast, later processes will be encountered, where one particle can be transmuted into another one by emitting a third particle. In these cases the relevant variables to be tuned are differently, either in a so-called $t = (p_1 - q_1)^2$ channel or, if the final particles are exchanged, in the $u = (p_1 - q_2)^2$ channel. These three variables together, s, t, and u, are denoted collectively as Mandelstam variables. Observing a resonance in any of the channels is already providing insights into its production process, as the different kinematics can be shown to stem from different physical origins.

Of course, in nature there is not a real divergence in the cross-section, since any resonances has its finite, though possibly small, width Γ. Including this width leads to a Breit-Wigner cross section in, e. g., the s-channel

$$\sigma \sim \frac{\Gamma^2}{(\sqrt{s} - m)^2 + \frac{\Gamma^2}{4}}. \tag{2.8}$$

Thus, the divergences becomes a peak with a width determined by the decay width of the particle, which is therefore also accessible to an experimental measurement.

Of course, this is an idealized situation. Higher order processes or processes of a different origin can all contribute to the cross-section. Already a second resonance close by could distort the cross-section appreciably. Furthermore, quantum mechanically interference processes could suppress a peak or, even worse, create a fake peak. This has been observed, and is therefore not just an academic possibility.

Finally, if the process is inelastic, there could be more than just two particles in the final state. Then the cross-section becomes more involved, and the presence of resonances is not just a simple peak. There are advanced techniques to deal with such situations, like the Dalitz analysis, which lead beyond the present scope.

One should note that the formula (2.8) is only accurate if Γ/m is small. For resonances with a large width, the shape can be significantly distorted to the point where no real peak is observable at all. However, if it is small, it is possible to connect it to the phase shift introduced in (2.5). To leading order, and close to the resonance $s \approx m$, the connection is given by

$$\cot \delta(s) = \frac{m^2 - s}{\Gamma\sqrt{s}}.$$

This follows directly from the combination of (2.6) and (2.8) for the scalar case $l = 0$. Especially, the presence of a resonance at $s = m$ is signaled by the phase shift going through π.

2.7 Feynman diagrams

As has now been repeatedly emphasized, scattering experiments are the central experimental tool in particle physics. Accordingly, theoretical calculations of them have become the single most important technique. The usual task is therefore to have some initial state, e. g. two protons at the LHC, and then determine the partial cross-section for a given final state, inclusive or exclusive.

The most well-known tool for doing so is perturbation theory, in which the interactions between the particles is expanded in a power series of its strength. The lowest order is usually called, under the assumption of quick convergence of the series, leading order, abbreviated by LO. The next order is called next-to-leading order, NLO, and so on as N...NLO or N^{n}LO. There are very many experimentally accessible processes in which already low orders in perturbation theory, usually one to four, yield the dominant part of the relevant cross-section.

However such an expansion is not able to capture all features of essentially all quantum field theories. An intuitive example, which is structurally similar to quantum field theories, is calculating the integral

$$\int_0^\infty dx e^{-x^2-\lambda x^4} = e^{\frac{1}{8\lambda}} \frac{K_{\frac{1}{4}}\left(\frac{1}{8\lambda}\right)}{4\sqrt{\lambda}},$$

where K_n is the modified Bessel function of the second kind, and $\lambda > 0$. This is a well-defined integral. Assuming $\lambda \ll 1$, it appears reasonable that this result should also be obtainable by expanding the integrand perturbatively into a power-series in λ, yielding

$$\int_0^\infty dx e^{-x^2-\lambda x^4} = \sum_n \int_0^\infty dx e^{-x^2} \frac{(-1)^n \lambda^n x^4 n}{n!} = \sum_n \frac{(-1)^n \lambda^n \Gamma\left(\frac{1}{2}+2n\right)}{2(n!)}.$$

However, as the Γ-function behaves like $n!$ for large n, this series diverges, showing that even for arbitrary small λ an expansion does not work. In fact, the series elements start to grow roughly from $1/n \sim \lambda$ onwards.

Though this is a typical situation in quantum field theories, this by no means implies that perturbation theory is useless. If performed only up to an order from which on the series elements start to grow, it is easily, and in fact very often, possible that quantitatively this already gives the bulk of the result. In fact, for most aspects of particle physics, perturbation theory suffices to be competitive to the experimental precision.

But there are cases, in which this is not the case. Most prominent among them is bound state physics - since perturbation theory starts usually around a non-interacting theory, it knows only about elementary particles, but not about (stable) bound states. It is therefore very advisable to analyze a problem before choosing a particular method to deal with it.

Since for much of the remainder of this lecture perturbation theory is indeed providing already the dominant contribution, is is worthwhile to introduce the language of perturbation theory. Furthermore, non-perturbative physics is almost always a generalization of perturbation theory. Hence, first analyzing a problem in a perturbative language and then generalizing is in most cases the wisest course of action.

Perturbative cross-section calculations in quantum field theories can be organized similar to quantum mechanics, i. e. as a mathematical series. Such a series can be graphically represented, which is called Feynman diagrams. In such diagrams a line corresponds to the propagation of a particle, a so-called propagator, while a vertex corresponds to an interaction. Thus, the number of vertices corresponds directly to the order of the series expansion. It can then be shown that each graph can be uniquely mapped to a mathematical expression describing the contribution of a given physical process to a cross-section. Thus one can use the mathematical tool of graph theory to construct all contributions of a given order to a physical process. An interesting observation is that the LO contribution is always a graph without loops, and therefore LO is also called tree-level. On the other hand, all higher orders have at most as many loops as the order, i. e. NLO at most one loop, and thus the expansion is also called loop expansion. Finally, it can be shown that every loop introduces one power of $\hbar$. Hence, the expansion is also an expansion in quantum effects, especially as at tree-level there is no non-trivial power of $\hbar$. Hence, tree-level is also called the classical contribution.

The possible vertices and propagators are not arbitrary, but are fixed for each theory in the form of so-called Feynman rules, which uniquely determine all the possible lines, vertices, and composition rules. These have to be determined for each theory anew, though this is an algorithmic procedure which a computer algebra system can perform. Especially vertices are restricted by conservation laws, like four-momentum conservation.

In practice for a reasonable complicated theory the number of diagrams grows factorial with the order of the expansion. Thus higher-order calculations require sophisticated methods to deal with the logistical problems involved.

Finally, it has turned out that also beyond perturbation theory a graphical language can very often be introduced in a mathematical precise way. Hence, Feynman-graph-like representations are ubiquitous in particle physics, though may mean very different mathematical objects. Especially, they are not necessarily restricted to a series expansion, and it is possible to encode also the full content of a theory graphically. Thus, one should always make sure what entities a given set of graphs corresponds to.

Chapter 3

A role model: Quantum electrodynamics

The first example of a particle physics theory, the first part of the modern standard model of particle physics, to emerge was quantum electrodynamics, QED. It already contains many of the pertinent features of the more complex standard model, while at the same time is much simpler. It is therefore the ideal starting point to get acquainted with a particle physics theory.

Quantum electrodynamics is describing the existence and interaction of electromagnetically active particles. As such, it is the quantum generalization of classical electrodynamics, which already implements special relativity.

3.1 Electrons and photons

QED, like any other particle physics theory, implements two roles for particles, though particles may later also implement both roles simultaneously.

The first role is that of charge carrier. Classically, these are objects which carry an electromagnetic charge. They therefore constitute the quanta of the electric current $j = (\rho, \vec{j})$. The prime example of these charge carriers are the electrons. These elementary particles are fermions, having spin 1/2, and have a rest mass of $m \approx 511$ keV. To the best of our current knowledge they are point-like[1], or at least smaller than 10^{-7} fm. Each electron carries exactly one unit of electric charge, which is in natural units dimensionless, and of a size of roughly $e = -0.3$. Its sign is chosen by convention, giving electrons a negative charge, in accordance with classical physics.

[1]The problem of infinite energy of point-like objects is resolved in quantum field theory, at least in an effective way. This will be discussed below in section 3.7.

The second role is that of the force carrier. Such particles are exchanged by charge carriers to transmit the force. The force carriers are thus the quanta of the electromagnetic field, which transmit electromagnetic interactions classically. Thus, the particles carrying the electromagnetic force are the quanta of the electromagnetic field, called photons. These are massless[2] bosons of spin one, and therefore so-called vector bosons.

In classical electrodynamics it is the electromagnetic fields $\vec{E}$ and $\vec{B}$, which are associated with the force. These fields can be derived from the vector potential A,

$$\begin{aligned} B_i &= \epsilon_{ijk}\partial_j A_k \\ E_i &= -\partial_i A_0 - \partial_0 A_i. \end{aligned}$$

Classically, the vector potential does not appear in the Maxwell equations

$$\begin{aligned} \vec{\nabla}\vec{E}(\vec{x}) = &\quad \partial_i E_i &= 4\pi\rho(\vec{x}) \\ \vec{\nabla}\times\vec{B}(\vec{x}) - \partial_t\vec{E} = &\quad \epsilon_{ijk}\partial_j B_k(\vec{x})\vec{e}_i - \partial_t\vec{E} &= 4\pi\vec{j}(\vec{x}) \\ \vec{\nabla}\times\vec{E}(\vec{x}) + \partial_t\vec{B}(\vec{x}) = &\quad \epsilon_{ijk}\partial_j E_k(\vec{x})\vec{e}_i + \partial_t\vec{B}(\vec{x}) &= 0 \\ \vec{\nabla}\vec{B}(\vec{x}) = &\quad \partial_i B_i &= 0 \\ \vec{\nabla}\vec{j} + \partial_t\rho = &\quad \partial_i j_i + \partial_t\rho &= 0, \end{aligned}$$

which describe the interaction between the electromagnetic fields and the sources. However, already in the quantum-mechanical Hamilton operator of a (spinless) electron

$$H = \frac{1}{2m}(i\partial_i - eA_i)^2 + eA_0, \tag{3.1}$$

it is instead the vector potential that couples to the particle. Thus, the photons are actually the quanta of the vector potential A, instead of the E and B fields[3]. Classically, for any given $\vec{E}$ and $\vec{B}$ fields the corresponding vector potential is not uniquely determined. It is always possible to perform a gauge transformation

$$A_\mu \to A_\mu + \partial_\mu\phi, \tag{3.2}$$

[2]The concept of mass in quantum field theory is a non-trivial one. What precisely is meant by the statement that the photon has zero mass involves a significant amount of technical subtleties, which will be glossed over here. A full discussion requires a treatment of quantum field theory beyond perturbation theory.

[3]It is actually possible to write down a quantum theory also in terms of the $\vec{E}$ and $\vec{B}$ fields. However, such a theory is much more complicated, especially it contains non-local interactions, and is almost intractable analytically. Hence the formulation in terms of the vector potential is used essentially exclusively. It is called the localized version, since no non-local interactions are present. However, this shows that the physical degrees of freedom are the electromagnetic fields, rather than the vector potential. Hence, using the vector potential is just a convenience.

with some arbitrary function ϕ. This is also true quantum-mechanically, except that also the wave-function ψ of the electron needs to be modified as

$$\psi \rightarrow \exp(-ie\phi)\psi. \tag{3.3}$$

Thus, the quanta describing electrons and photons are not uniquely defined, but can be altered. It can, however, be shown than any experimentally observable consequence of the theory does not depend on the the function ϕ, and thus on the choice of gauge. Theories with these features are called gauge theories. All experimentally supported theories contain at least one gauge field. They therefore represent the archetype of particle physics theories. QED is just the simplest example for such a gauge theory. Later on, much more complicated gauge theories will appear, which eventually form the standard model of particle physics.

Since the gauge symmetry requires that the ordinary derivative in the Hamilton (3.1) is replaced by the combination of the ordinary derivative, ∂, and the gauge field A_i, $D_i = \partial_i - eA_i$, this combination is called the covariant derivative. The name covariant stems from the fact that the whole operator transforms in such a way under a gauge transformation, as to make the whole Hamiltonian gauge-invariant[4]. This procedure is also known as minimal coupling. In principle, more complicated possibilities exist to obtain a gauge theory. But since so far no experimental support for such theories has been found, they will be skipped here.

It is worthwhile to mention that the gauge symmetry is essentially what was called a redundant variable in classical mechanics. In classical mechanics, the usual first step is to eliminate all redundant variables by enforcing all constraints. This is not done in particle physics, and the presence of gauge transformations (3.2-3.3) is just a manifestation of this redundancy. Keeping this redundancy is, in contrast to classical mechanics, technically much more advantageous than eliminating it. The presence of this gauge degree of freedom is thus rather a mathematical convenience, than a genuine feature of nature.

As noted before in section 1.3.4, there are anti-particles to every particle, and especially the positron as an anti-particle to the electron. It is thus also a spin 1/2 fermion with the same mass, but opposite electric charge $e = 0.3$. The photon is uncharged. It is therefore its own anti-particle, and there is no anti-photon which could be distinguished by any means from the photon.

[4]Note here the distinction between gauge-invariance, i. e. unchanged under a gauge transformation, and gauge-independent, i. e. not transforming at all under a gauge transformation.

3.2 The Dirac equation

As has been emphasized, the electron is a fermion. However, (3.1) only describes a scalar. The description of fermions in quantum field theory is somewhat more complicated. Hence, here only some of the most pertinent features will be given, and the remainder is left to a quantum field theory course.

The basic features distinguishing fermions from bosons is their half-integer spin, especially 1/2, and the Pauli principle. Half-integer spin is impossible for ordinary rotations. However, the Lorentz group permits the construction of such particles. The most important property is that it is not possible to write down a one-component wave-function for this purpose. At least, two (complex) components are needed. If a particle with an anti-particle should be described, as will be required exclusively for the standard model, two more components for this anti-particle are necessary. Thus, the simplest wave-function for QED is a four-dimensional complex object, a so-called spinor ψ. The name indicates that this quantity is different from an ordinary four-vector, despite having the same number of components. Especially, it does not transform in the same way as a four-vector under Lorentz transformations.

Without proof, how such a spinor transforms can be obtained from a ordinary Lorentz transformation. An ordinary Lorentz transformation acts on a four-vector v^μ as

$$V_\alpha = \left(e^{-\frac{i}{2}\omega_{\mu\nu}J^{\mu\nu}}\right)_{\alpha\beta} V^\beta,$$

where the $J_{\mu\nu}$ are the generators of boosts and rotations and the anti-symmetric matrix $\omega_{\mu\nu}$ contains the six free parameters, three angels and three boost levels, of a general proper Lorentz transformation. Note that this a matrix exponential, as the $J_{\mu\nu}$ are Lorentz tensors, and hence α and β sum over these Lorentz indices. Spinors transform then as

$$\Psi_\alpha = \left(e^{-\frac{i}{2}\omega_{\mu\nu}S^{\mu\nu}}\right)_{\alpha\beta} \Psi^\beta,$$

where just the generators are different. They are also 4×4 matrices, given by

$$S_{\mu\nu} = \frac{i}{4}\left[\gamma^\mu, \gamma^\nu\right],$$

where the four Dirac-γ matrices are any set of 4×4 matrices which satisfy the so-called Clifford or Dirac algebra

$$\{\gamma^\mu, \gamma^\nu\}_{\alpha\beta} = 2g^{\mu\nu}\delta_{\alpha\beta}.$$

The different possibilities just correspond to unitarily equivalent conventions. Note that the index μ is not implying that the γ-matrices are a four-vector. They are the same in any frame, and invariant under Lorentz transformations. Nonetheless, the metric can be used to define γ-matrices with raised indices as well.

A, for many purposes convenient, choice of the γ matrices is the so-called Weyl or chiral representation,

$$\gamma^0 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$
$$\gamma^i = \begin{pmatrix} 0 & \tau^i \\ -\tau^i & 0 \end{pmatrix},$$

where the τ^i are the ordinary Pauli matrices. However, the explicit form of the γ-matrices, as well as their properties, will not be needed throughout the lecture.

While this gives an object which transforms correctly as a spin 1/2 object, this is not yet providing a dynamical description for it. This is given by the postulate of the Dirac equation to describe a free fermion of mass m by a spinor ψ,

$$(i\gamma^\mu\partial_\mu - m)\psi = 0. \tag{3.4}$$

Note that the product $\gamma^\mu\partial_\mu$ is not a scalar, since the γ_μ are not a four-vector. This operator transforms covariantly under Lorentz transformation. Though the Dirac equation is a postulate, it has been confirmed experimentally, and is a viable description of a free fermion. Applying the Dirac operator, i. e. $i\gamma^\mu\partial_\mu - m$, on the Dirac equation yields

$$(\partial^2 + m^2)\psi = 0,$$

i. e. each component of the spinor fulfills the Klein-Gordon equation. Therefore, the relativistic relation between energy and momentum is obeyed also by fermions, making them reasonable particles.

The Dirac equation (3.4) can be obtained as the Euler-Lagrange equations from the Lagrangian

$$\mathcal{L} = \bar{\psi}(i\gamma^\mu\partial_\mu - m)\psi, \tag{3.5}$$

where $\bar{\psi} = \psi^\dagger\gamma_0$, and is the spinor of an anti-particle. This definition ensures that the Lagrangian is a scalar. Hence, it follows that the quantity $\bar{\psi}\psi$ is a scalar and $\bar{\psi}\gamma_\mu\psi$ is a vector under Lorentz transformations. It is furthermore possible to define the matrix $\gamma_5 = -i\gamma_0\gamma_1\gamma_2\gamma_3$, which can be used to create a pseudo scalar $\bar{\psi}\gamma_5\psi$ and an axial vector $\bar{\psi}\gamma_5\gamma_\mu\psi$. Furthermore, the relation between fermions and anti-fermions require that a mass term includes both of them, and there are no separate mass terms for both particle species. Especially, this implies that the fermion and the antifermion have the same mass.

The Pauli principle can be ensured by replacing the complex numbers of the spinor components by so-called Grassmann numbers. A Grassmann number ξ has the property that it squares to zero, i. e. $\xi^2 = 0$, and anti-commutes with any other Grassmann number η, $\xi\eta = -\eta\xi$. Thus spinor components anti-commute. This implies that the square of each of the spinor components vanish, if evaluated at the same space-time point, thus implementing the Pauli principle. Though this has substantial implications in quantum field theories, for the present purpose it is enough to just keep the anticommutativity of spinor components in mind.

3.3 Formulation as a field theory

So far, the formulation at the level of (3.1) is quantum-mechanical. As noted, quantum mechanics cannot deal with relativistic effects, and one has to pass to quantum field theory (QFT). QFT is technically much more complicated than ordinary quantum mechanics, and is beyond the scope of this lecture. However, it is very useful to understand and use the way QFTs are formulated, and this provides a quick way of representing a theory. Furthermore, significant insight can be gained already by treating the QFT classically, and thus just like classical electrodynamics, which is a classical field theory.

This should now be exemplified for the case of QED.

The simplest part is just starting with Maxwell theory, i. e. the theory of free electromagnetic fields. Since the Hamilton operator is not a Lorentz-invariant it is usually not the best way to formulate a relativistic theory. A better starting point is the Lagrange function L. For a field theory like electromagnetism, this changes again to a Lagrangian density $\mathcal{L}$, the Lagrangian. Classical Maxwell theory takes the form

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} \tag{3.6}$$

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu, \tag{3.7}$$

which is explicitly invariant under the gauge transformation (3.2). The anti-symmetric tensor $F_{\mu\nu}$ is the field-strength tensor, with components formed by the electromagnetic fields. The fact that the photon is a spin one boson can be read off from the vector potential since it carries a single Lorentz index, and therefore transforms like a vector, and is described by an ordinary field.

The total Lagrangian of QED is obtained by coupling electrons and positrons to the electromagnetic field. As in the quantum-mechanical case, this is achieved, and experimentally confirmed, by minimal coupling. Because of the fermionic nature of electrons, this is applied to the Dirac Lagrangian (3.5), yielding

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \bar{\psi}\gamma_\mu(i\partial^\mu - ieA^\mu)\psi - m\bar{\psi}\psi, \tag{3.8}$$

where m is the mass of the electron. The spinors representing the electron transform under gauge-transformations like the quantum-mechanical wave-functions by the phase (3.3), and therefore this Lagrangian becomes gauge-invariant in the same way as in the quantum-mechanical case.

Besides the presence of photons and (massive) electrons and positrons another property of the theory can be read off the Lagrangian (3.8). There is a term which involves not two, but three fields, the interaction term $ie\bar{\psi}\gamma_\mu A^\mu\psi$. This is the first example of an interaction vertex, which specifies an interaction of three particles, an electron, a positron, and a photon. They interact with the coupling strength e, the electromagnetic coupling. The remainder factor $i\gamma_\mu$ ensures that the vertex respects Lorentz invariance, and that it yields a meaningful quantum theory. Note that there is no term describing interactions with more than three particles in this Lagrangian. This will change later on.

It is a further remarkable feature that the involved fields are functions not only of space, but also of time. Thus, a field configuration, i. e. any given field, represents a complete space-time history of a universe containing only QED.

As emphasized before, the Lagrangian is a classical object, i. e. the fields are not operator-valued. The quantization yields a QFT, a topic treated in a separate lecture. Here, it suffices to use the language of Lagrangians to specify a theory.

3.4 Gauge symmetry

The gauge transformations (3.2-3.3) have indeed a quite deeper structure than visible at first sight. Especially, they implement a symmetry, the gauge symmetry. The electron fields are modified by an arbitrary phase factor at every space-time point. This phase factor is a complex number, and therefore corresponds to a two-dimensional rotation. It is therefore an explicit implementation of the two-dimensional rotation group SO(2). Since the two-dimensional plane is equivalent to the complex plane, two-dimensional rotations can also be seen as phase rotations, and the group of all phase rotations is called U(1). This equivalence is also expressed as SO(2)$\approx$U(1). Since phase factors are more useful in quantum physics, the latter formulation is usually employed. As a consequence, the gauge symmetry is called a U(1) gauge symmetry, and the group U(1) is called the gauge group. Since the elements of U(1) commute, this is also called an Abelian gauge theory.

The electron field transforms like an element of this group, i. e. by multiplication, (3.3). The electron is therefore said to be in a representation of the gauge group, the so-called fundamental representation.

The photon field does not transform by a phase factor, but instead by an additive function, (3.2), which is related to the phase factor by exponentiation. It can be shown that these functions implement an algebra, the so(2)≈u(1) algebra. Therefore, the photon field is said to be in a different representation of the gauge group, the so-called adjoint representation.

Such structures will again reappear for the other parts of the standard model, and are also present in gravity. So far, only the fundamental and adjoint representation play a role in particle physics, though several proposals exist for theories beyond the standard model in which particles in different representations would appear. They would then transform differently under gauge transformations. The precise form is given by the relevant gauge group, and is mainly an exercise in group theory. A more formal discussion of this will be given below in section 4.1.

For comparison, gravity in the form of the general theory of relativity can also be thought of as a gauge theory, with as gauge group T^4, where T is the (Abelian) translation group in one direction. This is, however, beyond the scope of this lecture.

3.5 Bound states and the positronium

Already the simple system of QED in form of (3.8) yields a plethora of interesting phenomena, which will be discussed in the following. The first one is that besides the electrons and photons[5] it is possible to construct bound states. The simplest bound state usually encountered is an atom. Lacking a proton or nuclei so far, the simplest bound state, which can be constructed in QED is positronium. This bound state consists out of an electron and a positron. In principle, it is just like a hydrogen atom, but its nuclei is just as heavy as the electron. This state has also been observed in experiments.

Of course, such a bound state is not stable. As stated, matter and anti-matter can annihilate each other, and so can electrons and positrons. Thus, after a while, the electron and the positron will annihilate each other into two or three photons (one is not permitted due to four-momentum conservation). The two different options depend on the relative alignment of the electron's and positron's spins, either parallel or anti-parallel. Since the life-time of these states is substantially different, they are also called differently,

[5] Again here some subtleties arise in a quantum-field theory when one talks about electrons or photons as particles. This is also beyond the scope of this lecture.

orthopositronium and parapositronium.

Thus, even such a simple theory as QED has already highly non-trivial structures.

3.6 Muons and Taus

Before delving further into the theoretical features of QED, it is time to introduce a property of nature which was not foreseen at the time QED was invented, and for which till today no convincing explanation has been found. This is the fact that the electron has two heavier siblings, and likewise does the positron. One is the muon, having a mass of about 105 MeV, and the other the tauon, with a mass of around 1777 MeV. Thus, both are much heavier than the electron. Otherwise their properties are the same as for the electron: They are fermions with spin 1/2 and negative electric charge. They can be easily included into the Lagrangian (3.8),

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \sum_f \bar{\psi}_f \gamma_\mu (i\partial^\mu - ieA^\mu - m_f)\psi_f, \tag{3.9}$$

where f now counts the so-called flavor, i. e. type, of the fermions, and can be electron, muon, and tauon. In QED, this flavor is conserved, i. e. it is not possible to convert the heavier muon into an electron. Later, additional interactions will be encountered, which change this. Then neither the muon nor the tauon are stable, and thus are encountered in nature only as short-lived products of some reaction.

The electron, muon, and tauon are otherwise exact replicas of each other. They form the first members of three so-called families or generations of elementary particles, which will have, except for the mass, otherwise identical properties. Their presence, and the relative sizes of their masses form one of the most challenging mysteries of modern particle physics, subsumed under the name of flavor physics. However, several interesting phenomena we observe in nature are only possible if there are at least three such families. Some will be encountered later on.

To distinguish the electron, muon, and tauon from other particles to be discussed later, they are also collectively denoted as leptons.

3.7 Radiative corrections and renormalization

When one tries to perform perturbative calculations in QED, this is actually rather straightforward to leading order. However, a serious problem occurs if an attempt is made to go beyond tree-level. These were first encountered when people attempted to calculate corrections due to the emission and absorption of photons from electrons bound in atoms. Therefore, such contributions are nowadays collectively called radiative corrections, though in most cases this name is less than accurate.

The problem encountered in the calculation stems from the combination of two features of quantum field theories. One is that energy may be borrowed for some times in the form of virtual particles. The other is that quantum mechanically everything what could happen happens and mixes.

Therefore, in the next order of perturbation theory, NLO, generically the possibility appears that an electron can emit a virtual photon and then reabsorb it later. Or a photon can split into a virtual electron-positron pair, which later annihilates to form again the photon. So far, this is not a problem. But these virtual particles can borrow energy, and therefore may be created with arbitrarily large energies. These arbitrarily large energies then induce divergences in the mathematical description, and all results become either zero or infinite.

This is, of course, at first sight a catastrophe, and for a while it seemed to mark the end of quantum field theory. It is actually not an artifact of perturbation theory, as one might suspect; though there are theories where this is the case, this is not so for those relevant for the standard model of particle physics[6].

Today, this failure has been understood not as a problem, but rather as a signal of the theory itself that it cannot work at arbitrarily high energies. It signals its own breakdown, and the theory is, in fact, only a low-energy approximation, or a low-energy effective theory, of the true theory. This is then as expected, since gravity is not yet part of the standard model, but is expected to play an important role at very high energies. Hence the prediction of failure is consistent with our experimental knowledge on the presence of gravity.

This insight alone does not cure the problems, however. Fortunately, it turns out that a certain class of theories exist, so-called renormalizable theories, where this problem can be cured in exchange for having a finite number of undetermined parameters. In case of the full QED Lagrangian (3.9) these are four: The strength of the electromagnetic coupling and the three masses. We exchange the divergences in favor of having these four numbers as an input to our theory, which we have to measure in an experiment. Once these parameters are fixed, reliable predictions at not too high energies can be made, where not too high for the standard model covers an energy range larger than our current experimental reach, probably much larger.

[6]There are a few special theories where this is not the case, especially in dimensions lower than 4. However, none of them is part of the experimentally established particle physics.

This exchange, a process known for historical reasons as renormalization, was the final savior of quantum field theory, making it again a viable theory. It also implies that a renormalizable theory cannot be a complete description of physics: The parameters of the theory cannot be predicted from inside the theory, and external input is needed. Whether a completely parameter-free theory can be formulated, which describes nature, is not known.

What is known is that whether a theory requires only a finite number of parameters can in most cases be deduced by simply looking at the Lagrangian. This is known as superficial power-counting. Since the action has to be dimensionless in natural units, but the integral as well as derivatives have fixed dimensions, this implies that the fields themselves have to have dimension. From looking at the kinetic terms, i. e. those with only two fields, it follows immediately that scalars and gauge fields have dimension of energy, while fermions have dimension of energy$^{3/2}$. Furthermore, mass has, as required, dimension of energy. This implies that the electromagnetic coupling is dimensionless, as it ought to be. It can then be shown that, in most cases, a theory requires only a finite number of parameters if all numbers, masses and coupling constants, in the Lagrangian have zero or positive dimension in terms of energy. Hence, QED is renormalizable. The notable exceptions to the rule will be encountered later. E. g., the Higgs theory was invented because without the Higgs particle the standard model would be non-renormalizable despite being superficially renormalizable. On the other hand there exist some theories, possibly including quantum gravity, which are superficially non-renormalizable, but in fact are renormalizable.

3.8 Fixing the inputs and running quantities

In the previous section 3.7, it was sloppily said that it is necessary to fix the parameters, in case of QED three masses and one coupling. This is a far less trivial process than it seems at first.

The problem is most easily seen when it comes to fixing the electric charge. To do so requires to define what is meant by the term electrical charge. Classically, this is simple. Together with the electron mass, it can be determined by the deflection of an electron in both an electric an a magnetic field. Quantum-mechanically, this is not so obvious. As stated, what is observed in nature is not just the electron. What actually looks like an electron, is an electron which is constantly emitting virtual photons, which themselves may emit virtual electron-positron pairs. It is this electromagnetic cloud, which surrounds the electron, what is measured. When the electron is probed at shorter and shorter distances, the probe travels farther and farther into this cloud. Therefore, what is measured, depends on the distance traversed, and thus the energy of the probe. Hence, it is necessary to make a definition of what electrical charge is, since it is not possible to measure only the process at tree-level.

For QED, this is usually done by measuring Thompson scattering. This is the scattering cross-section of a photon of small energy from an electron at rest. Aside from some trivial geometric and kinematic factors, the size of this scattering is determined by the fine-structure constant $\alpha = e^2/(4\pi)$, and thus by the electric charge. By fixing the parameters in the Lagrangian of the theory (3.9) such as to reproduce this scattering cross-section, one of the parameters is fixed, in this case to the value of roughly 1/137.

This at first sight simple prescription has two inherent subtleties. The first is rather of a technical nature. For no relevant theory it is possible to calculate things like a scattering cross-section exactly. As a consequence, the dependence of the calculated cross-section on the parameters of the Lagrangian is only approximate, and therefore their determination can only be approximate. Moreover, two different calculations done with different approximations will not necessarily yield coinciding values for the parameters. It is therefore mandatory to make sure that any comparison has to be done in such a way as to take such approximation artifacts into account.

The second subtlety is far more complicated. It has just been defined what the electric charge is. However, it is not forbidden to instead define the electric charge by measuring the scattering cross-section for an electron and photons having all a certain energy, say μ. This is the so-called symmetric configuration. In general[7], the scattering cross-section will depend on this energy μ, and hence the so-defined electric charge will be different, and depend on the actual value of μ, $\alpha = \alpha(\mu)$. Thus, there is no unique definition of the electric coupling, but a coupling depending on the energy where it is measured. This is called a running coupling, a concept of great importance.

The change of the running coupling with the energy scale μ is also called the renormalization evolution, and it is possible to write down a mathematical group where the elements are changes in this scale. This is the so-called renormalization group, a very powerful tool when performing actual calculations in quantum field theory.

Similarly, the electromagnetic cloud of an electron will be deformed, if it is probed at higher and higher energies. As a consequence, the effective mass of an electron is also not independent of the energy[8], and the mass depends similarly on the energy $m_e(\mu)$, a running mass.

This running of quantities is a characteristic of quantum field theories. In fact, testing

[7]There are some theories, so-called conformal theories, where this is not the case. None of the theories realized in nature are of this type, though. This can actually be excluded exactly from experiment.

[8]Here, again, a subtlety of advanced quantum field theory plays a role when it comes to define what an electron is. This will also be skipped, as it is of little effective importance.

the energy-dependence of the running quantities is one of the most stringent tests, which can be performed on the validity of a theory. For the standard model, this has been done for the running over several orders of magnitude, a marked success of the theory.

It now may appear that one parameter has been traded in for a function. This is not the case. It is actually sufficient to measure the value of four running quantities at a single energy to fix all parameters of QED uniquely. The remainder of the functions is then uniquely determined by the theory and the four parameters. It is thus sufficient to probe a theory at a single energy scale, and it is then possible to make predictions about it at completely different energy scales.

A final word of caution must be added. In the present case, the electric charge was defined at the symmetric point, to obtain a single running coupling. Usually, this choice is also normalized such that in the absence of quantum corrections the classical coupling emerges, as done here when taking the Thomson limit. There is no reason to make this choice. An equally well acceptable choice would be to chose the photon's energy twice as large as the one of the electron, and deduce everything from this configuration. In the end, physics must be independent of such choices. This is indeed what happens.

Such different choices are called renormalization schemes, and the change between two such schemes is called a renormalization scheme transformation. Though a mathematical well-defined process, it is important to compare only results in the same scheme, as otherwise the comparison is meaningless.

3.9 Landau poles

One of the mainstays of modern particle physics remains perturbation theory. However, as noted before, perturbative calculations can, strictly speaking, never be entirely correct, and there are always non-perturbative contributions to any quantities at all energies. Still, in many relevant cases, the perturbatively calculable contribution is by far the dominant part.

One of the arguably most famous results of perturbation theory is the aforementioned running of the coupling $\alpha(\mu)$. It shows that the running coupling satisfies the differential equation

$$\frac{de}{d\ln\mu} = \beta(e) = -\beta_0 \frac{e^3}{16\pi^2} + \mathcal{O}(e^5), \tag{3.10}$$

where β is the so-called β-function, and β_i are its Taylor coefficients. The latter can be calculated in perturbation theory. In fact, the value of the coefficients are known up to order 10 in α, i. e. 20 in e, in the expansion. Truncating the series at the lowest order, the differential equation can be integrated, and yields

$$\alpha(\mu^2) = \frac{e(\mu^2)^2}{4\pi} = \frac{\alpha(\mu_0^2)}{1 + \frac{\alpha(\mu_0^2)}{4\pi}\beta_0 \ln \frac{\mu^2}{\mu_0^2}} \equiv \frac{4\pi}{\beta_0 \ln \frac{\mu^2}{\Lambda^2}}, \tag{3.11}$$

where μ_0 is the energy where the experimental input $\alpha(\mu_0)$ has been obtained. The so introduced quantity Λ is often called the scale of the theory, and is a characteristic energy of the described interaction. The value of β_0 for QED is -4. This implies that the running coupling becomes larger with increasing energy, until it eventually diverges at a very large energy scale.

This divergence is the so-called Landau pole. Similar problems arise also in the other running quantities. The presence of such a Landau pole can be traced back to the use of perturbation theory. Before reaching the Landau pole, the running coupling, which is the expansion parameter, becomes larger than one, and therefore the perturbative expansion breaks down. Thus, a Landau pole is a signal of the breakdown of perturbation theory.

When this point is reached, it is necessary to resort to non-perturbative methods. Even for QED such a full non-perturbative treatment is not simple. The results, however, indicate that QED is also breaking down beyond perturbation theory, and the only quantum version of QED which make sense is the one with $e = 0$, a so-called trivial theory. This is the triviality problem. It is assumed that this effect is cured, once QED is embedded into a larger theory, in this case the standard model of particle physics. This is a recurring problem, and will be discussed in more detail later.

Chapter 4

Invariances and quantum numbers

4.1 Symmetries, groups, and algebras

As has already been seen in section 3.4, symmetries and groups play an important role in particle physics. This becomes even more pronounced when eventually the whole standard model will be formulated. A technically convenient and tractable formulation of particle physics appears so far only possible by using gauge symmetries. However, these are not the only kind of symmetries which play an important role in particle physics. The most important ones will be discussed in the following sections.

Invariances have two important consequences. One is that the observation of an invariance strongly restricts how the underlying theory can look like: The theory must implement the invariance. The second is that if such an invariance is implemented, then as a consequence different processes can be connected with each other. E. g., one of the invariances to be encountered below is connected with the electric charge. As a consequence the scattering of two particles with given charges will have the same cross section as for the anti-particles with electric charge reversed. Hence, invariances are powerful tools in theoretical calculations, and powerful experimental tools to unravel the structure of the underlying theory.

An important classification of symmetries should be mentioned beforehand. In particle physics, there are three most important classes of symmetries. The first are discrete symmetries, like reflection in a mirror. They act, and do not involve any parameters. In contrast, continuous symmetries have transformations which depend on one or more parameters. If this parameter is constant, this is called a global symmetry. If it depends on space-time, as for the gauge transformations of QED, it is called a local or gauge symmetry.

Since symmetries imply the existence of an operator which commutes with the Hamilton operator, they lead in general to conserved quantum numbers. These can be either multiplicatively or additive. This means that a for a composite state the quantum numbers of the constituents are either multiplied or added. Discrete symmetries usually lead to multiplicative ones, while continuous ones to additive ones. Note that not all states have well-defined quantum numbers, even if a symmetry exists, and that a composite state can have well-defined quantum numbers even if the constituents do not have.

Furthermore, it is necessary to introduce a few notions of group theory. Later, in section 5.15, this will be further elaborated. For now, it is sufficient to introduce a few concepts. Most of these concepts are already familiar from quantum mechanics, especially from the theory of atomic spectra, and thus angular momenta. However, these concepts require generalization in the context of particle physics.

The basic property of a symmetry is that there exists some transformation t, which leaves physics invariant. Especially, such a transformation must leave therefore a state $|s\rangle$ invariant, up to a non-observable phase. Thus, a transformation has to be implemented by a unitary or anti-unitary operator, T. Especially in cases of continuous symmetries, it is possible to act on any state repeatedly with different values of the relevant parameter(s), $T(\alpha_1) = T_1$, $T(\alpha_2) = T_2$ yielding $T_2T_1|s\rangle$. There is also always a unit element, i. e. one transformation which leaves every state invariant. It is also furthermore for a unitary transformation possible to reverse it, T^{-1}. Hence, such transformations have a group structure, and it is said that they are the corresponding symmetry group.

An example has been the group of the gauge transformations from section 3.4. There, the group was SO(2)≈U(1). The group elements were the phase rotations $\exp(i\alpha(x))$, with space-time-dependent parameters. It was thus a local symmetry. The unit operator is thus just the one, with $\alpha = 0$. The inverse one is just $\exp(-i\alpha(x))$, the complex conjugated one, as expected for a unitary operator.

Since the application of two gauge transformations commute, this symmetry group is Abelian. However, a consecutive application of more than one group element need not be commutative,

$$T_1T_2|s\rangle \neq T_2T_1|s\rangle.$$

If it is not commutative, the symmetry group is called non-Abelian.

The gauge field did not transform in the same way, see equation (3.2). In fact, the transformation did not seem to be unitary at all. It is, however, only the manifestation of a different mathematical version of the symmetry. Since interesting symmetry transformations are (anti-)unitary, they can always be decomposed into a complex conjugation or a one, for anti-unitary or unitary transformations, respectively, and a unitary transformation. Any unitary transformation can be written as[1]

$$T(\alpha) = e^{i\alpha_i\tau^i}, \tag{4.1}$$

where the τ^i are Hermitian operators, and the possibility of a multi-parameter transformation has been included. These operator form the so-called generators of the symmetry group, and form themselves the associated symmetry algebra. Hence, the gauge fields transform not with the group, but with the algebra instead.

To separate both cases, it is said that the electron states and the gauge fields transform as group elements and as algebra elements, respectively.

A particular important special case arises, when the parameters α_i are (infinitesimally) small. The relation (4.1) can then be expanded to give the infinitesimal transformations

$$\begin{aligned} T &\approx 1 + i\alpha_i\tau^i \\ T^{-1} = T^\dagger &\approx 1 - i\alpha_i\tau^i. \end{aligned}$$

In many cases, it is sufficient to use such infinitesimal transformations to investigate the properties of a theory. However, this is not always possible, already in the standard model, though this will not be needed in this lecture. It just serves as a cautionary remark.

To illustrate the concept, the example of translational symmetry is useful. If a theory is invariant under translations, $x \to x + \Delta x$, a state changes under infinitesimal translations by

$$\psi(x+\Delta x) \approx \psi(x) + \Delta x\frac{d\psi(x)}{dx} = \psi(x) + i\Delta x\frac{d\psi(x)}{idx} = \psi(x) - i\Delta x\hat{p}\psi(x) \approx e^{-i\Delta x\hat{p}}\psi(x),$$

where $\hat{p}$ is the momentum operator. In this case, the momentum operator is the generator of translation, and the group elements are $e^{i\Delta x\hat{p}}$, forming the (Abelian) translation group. If the space is three-dimensional, the generators are the three spatial momenta $\hat{p}_i$, and the group elements become $\exp(i\Delta\vec{x}\cdot\hat{\vec{p}})$, a multi-parameter group.

4.2 Noether's theorem

One of the central reasons why symmetries are so important in quantum physics is that it can be shown that any continuous symmetry entails the existence of a conserved current. Such conserved quantities are of central importance for both theoretical and practical reasons. Practical, because exploiting conservation laws is very helpful in constructing experimental signatures which are not too contaminated by known physics or making

[1]Note that this relation is in general only unique up to a discrete group.

theoretical calculations more feasible. Theoretical, conserved currents mainly conserve quantum numbers. Since states of different conserved quantum numbers do not mix, they can be used to classify states. E. g. a state of positive charge will never mix with a state of negative charge in the sense that under time evolution an electron will not change into a positron.

The connection of symmetry and conserved currents is established by Noether's theorem. It is already a classical statement. Essentially it boils down to the fact that if the Lagrangian is invariant under a variation of the field, the variation can be used to derive a conserved current. To keep it simple, start with a Lagrangian $\mathcal{L}(\phi, \partial_\mu\phi)$ of a single real field ϕ, and its derivative $\partial_\mu\phi$. The Lagrangian is assumed to be invariant under the transformation

$$\phi \to \phi + \delta\phi,$$

i. e.

$$\mathcal{L}(\phi + \delta\phi, \partial_\mu\phi + \delta\partial_\mu\phi) = \mathcal{L}(\phi, \partial_\mu\phi) + \delta\mathcal{L} = \mathcal{L}(\phi, \partial_\mu\phi) + \partial_\mu K^\mu, \tag{4.2}$$

where $\partial_\mu K^\mu$ is a total divergence, which is neither classically nor quantum-field theoretically able to alter the behavior of the system[2]. The corresponding equation of motion of the field ϕ is the usual Euler-Lagrange equation

$$\partial_\mu \frac{\delta\mathcal{L}}{\delta\partial_\mu\phi} - \frac{\delta\mathcal{L}}{\delta\phi} = 0. \tag{4.3}$$

The explicit form of the variation (4.2) with respect to the field ϕ is then

$$\begin{aligned} \delta_\phi\mathcal{L} &= \delta\phi\frac{\delta\mathcal{L}}{\delta\phi} + (\delta\partial_\mu\phi)\frac{\delta\mathcal{L}}{\delta\partial_\mu\phi} \\ &= \delta\phi\partial_\mu\frac{\delta\mathcal{L}}{\delta\partial_\mu\phi} + (\delta\partial_\mu\phi)\frac{\delta\mathcal{L}}{\delta\partial_\mu\phi} \\ &= \partial_\mu\left(\delta\phi\frac{\delta\mathcal{L}}{\delta\partial_\mu\phi}\right), \end{aligned}$$

where the equation of motion (4.3) were used in the derivation and differentiations have been exchanged. This implies that the Noether current

$$j_\mu = K_\mu - \delta\phi\frac{\delta\mathcal{L}}{\delta\partial_\mu\phi},$$

or its corresponding extension to arbitrary sets of fields

$$j_\mu = K_\mu - \delta\phi_i\frac{\delta\mathcal{L}}{\delta\partial_\mu\phi_i} - \delta\phi_i^\dagger\frac{\delta\mathcal{L}}{\delta\partial_\mu\phi_i^\dagger}$$

[2] A least in flat space-time and for most quantum-field theories.

is conserved, $\partial_\mu j^\mu = 0$. The remaining of K^μ may seem a bit odd at first, as all variations of the field appears to be already included in the first term. K^μ includes all other terms, which may arise in the transformation without this change, e. g. by a change of boundary conditions.

Hence, the associated charge

$$Q = \int d^3\vec{x} j_0,$$

is conserved and thus time-independent,

$$\frac{dQ}{dt} = \int\limits_V d^3\vec{x}\partial_t j_0 = \int\limits_V d^3\vec{x}\vec{\nabla}\vec{j} = \int\limits_{\partial V} d^2\vec{n}\vec{j},$$

where $\vec{n}$ is the unit vector on the surface ∂V, stemming from the application of Gauss' law[3]. Thus, the only change in the total charge is by the migration of charge carriers through the surface in which the charge is enclosed. Hence, if the surface is send to infinity, where no more charges can come from the outside, the total charge becomes time-independent.

In this way, conserved currents are linked to symmetries of the theory. This will be encountered throughout the standard model of particle physics. Another useful feature of symmetries in quantum-field theories is that they entail relations between different quantities, so-called Ward-Takahashi identities for global symmetries and Slavnov-Taylor identities for local (gauge) symmetries. This is a rather technical topic in detail, and will therefore not be addressed further here.

4.3 Electric charge

It is worthwhile to apply this mathematical exercise to derive the electromagnetic current of (Q)ED. There are now three terms, the variations of the fermions, anti-fermions, and the photon-field. In this case, there appears no surface current K_μ. The three terms are

$$\begin{aligned} \delta\psi\frac{\delta\mathcal{L}}{\delta\partial_\mu\psi} &= (-ie\psi)\times(i\bar{\psi}\gamma_\mu) \\ \delta\psi^\dagger\frac{\delta\mathcal{L}}{\delta\partial_\mu\psi^\dagger} &= (ie\psi^\dagger)\times 0 \\ \delta A_\nu\frac{\delta\mathcal{L}}{\delta\partial_\mu A_\nu} &= (\partial_\nu)\times F_{\mu\nu} = 0, \end{aligned}$$

[3]Note that it was throughout tacitly assumed that the current and fields vanish sufficiently fast at infinity to permit the integral manipulations. In quantum-field theory, this is not obviously true anymore. But appropriate generalizations hold true in most cases, but not always, and care has to be taken. For the content of this lecture, this can be assumed to be true.

where the transformation function ϕ has been dropped, as it is arbitrary, and all identities have to hold anyway for any arbitrary choice. The product $\times$ symbolizes the adequate composition rules for the various tensors involved. Note that the term for the gauge field vanishes because of the antisymmetry of the field-strength tensor. Thus the current is

$$j_\mu = -e\bar{\psi}\gamma_\mu\psi, \tag{4.4}$$

where use has been made of the fact that the fermion fields anticommute, and there is an implicit γ_0 in the anti-fermion field. As expected, the current is carried only by the fermion fields, and the photons do not contribute. It thus reproduces the expected form.

Thus electric charge is conserved in QED, as it is in classical physics. This will also not be altered when passing to the standard model of particle physics. However, there is a remarkable feature. Since the gauge-transformation of the photon field (3.2) is independent of the electric charge, it is in principle possible that the three generations, i. e. electron, muon, and tauon, could have different and arbitrary electric charges. Furthermore, there is no reason that the proton has the opposite and positive charge as the electron, as it is not made of of positrons, as discussed below. Still, all experimental results agree to very good precision with this equality, especially the latter one. In QED, there is absolutely no reason for this fact. A discussed later, the standard model of particle physics is only a consistent quantum-field theory if and only if all the electric charges fulfill certain relations.

However, the fact that a theory only works if certain experimental facts are taken into account is supporting the theory. But is does not dictate that the experimental facts have to be this way. If a small deviation of these relation would be observed tomorrow, it is the theory which has the problem. Therefore, the condition that the electric charges are as they are for the theory to work is not an explanation of why this has to be. It is an experimental fact which can be described, but not explained inside the standard model. Though there are several proposals why this could be the case, especially so-called grand-unified theories. There is not yet any experimental support for any explanation, and it remains one of the great mysteries of modern particle physics.

4.4 Implications of space-time symmetries

Noether's theorem is not limited to so-called internal symmetries, i. e. symmetries which leave the space-time unchanged. It can also be applied to space-time transformations like translations and rotations. The consequences of both is that four-momentum and total angular momentum are conserved.

The spin of a particle is also a consequence of space-time symmetry, but in a much more subtle way. It is possible to show that the Poincare group only admits certain values of spin for particles, half-integer and integer[4]. This does not yet imply anything about the statistics which these particles obey. This requires further symmetries to be exploited later.

All this depends on having Minkowski space-time implementing the Poincare symmetry. These are two inputs to particle physics, and therefore the above listed properties are again not a prediction but a feature of particle physics. Other space-times can endow a theory with quite different properties, but for the subject of this lecture, it is fixed.

It should also be noted that the individual components of the spin have no meaning as conserved quantities in particle physics. This is most immediately visible when performing a Lorentz boost, under which they are not invariant. But for massless, and thus light-like particles, there is an exception. Since there is no way by making a Lorentz-boost to move into their rest-frame, the projection of the spin upon their momentum is always the same. This projection is called helicity, and is sometimes a useful concept in particle physics, when dealing with massless, or nearly massless, particles.

Note that space-time symmetries are global. Making them local leads to general relativity, which is not yet part of the standard model.

4.5 Parity

Another symmetry, closely related to space-time symmetry, is the discrete parity symmetry. The discrete transformation parity P is essentially a reflection in the mirror, i. e. $\vec{x} \to -\vec{x}$. Time is unaltered. A distinction between space and time appears here, because of the signature of the Minkowski metric, which makes time and space (in a loose sense) distinguishable. Note that applying a parity transformation twice yields again the original system. In fact, a parity transformation inverts the sign of each vector, e. g., coordinate r or momenta p

$$Pp = -p.$$

Pseudo-vectors or axial vectors, however, do not change sign under parity transformation. Such vectors are obtained from vectors, e. g., by forming a cross product. Thus the prime example of an axial vector are (all kind of) angular momenta

$$PL = P(r \times p) = Pr \times Pp = r \times p = L.$$

[4]In two space-time dimensions this is not the case, and the spin can take on any real number, leading to so-called anyonic particles. They are of no relevance in the standard model, but play a certain role in solid-state physics.

As a consequence, fields can either change sign or remain invariant under a parity transformation[5]. Hence, a quantity can be assigned a positive or a negative parity, $+1$ and -1. Generically, when there are only these two possibilities, the quantities are called even or odd, respectively, in this case under parity transformations.

The fields describing particles have definite transformation properties under parity. It is necessary to define the parity of some states, to remove some ambiguities regarding absolute phases. Thus, the absolute parity of a particle is a question of convention. Usually, the electrons, and thus also muons and tauons, are assigned positive parity, while the photon has negative parity. Note that the anti-particles have the opposite parity for fermions, but the same for bosons. The reason is the different statistics.

In QED, parity is conserved. Classically that can be read off the Lagrangian (3.9), though there are some subtleties involved concerning fermions when performing the transformation. Furthermore, it it is important to note that also differential operators have definite transformation properties under parity transformation, and this is the same as the one of the photon field. As a consequence all terms have a definite positive parity, and the total Lagrangian has even parity. In principle, this could be changed in the quantization procedure, but this does not occur for QED. Later, additional interactions of particle physics will break this symmetry already classically.

4.6 Time reversal

Of course, the corresponding partner to parity is the discrete time-reversal symmetry T, i. e. the exchange of $t \to -t$, without changes to the spatial coordinates. This is physically equivalent to reversing all speeds and momenta, and as a consequence also angular momenta: Either objects move backward or the time moves backward. Time-reversal symmetry is anti-unitary, and therefore involves besides a factor of 1 or -1 also a complex conjugation of the object it acts upon. Similarly to parity, fields can be classified as being even or odd under time-reversal. However, because time-reversal is anti-unitary, there is no quantum number associated with it.

As with parity, time reversal symmetry is respected both classically and quantum-mechanical by QED, and again later on theories will be encountered which do not respect it.

[5]For fermions, there is a multiplication by γ_0 involved.

4.7 Charge parity

In classical electrodynamics, it is possible to reverse the sign of all electric charges without changing the physics. This is a classical example of charge parity.

In particle physics, it is the presence of particles and anti-particles, and the fact that all their intrinsic (additive) properties are identical, that permits to define another discrete symmetry, the quantum version of charge parity C. Charge parity defines how a quantity behaves under the exchange of particles and anti-particles. Fields are also complex-conjugated, making charge-parity an anti-unitary operation. Note that properties like rest mass, momentum, spin and angular momentum are not affected by charge parity.

It is again possible for a system to either change sign or not, defining even and odd charge parity. But it is also possible to not have a particular charge parity at all, which is particularly true for charged states. That can be understood in the following way: To have a definite charge parity, a state must be an eigenstate under the charge parity transformation. However, charge parity transforms a particle with, say, electric charge +1 into a state with charge -1, and since such states are not identical, these states cannot have a definite charge parity. However, neutral states like the photon can have. Its charge parity, e. g., is -1. Also, states made up from multiple charged states which are in total neutral can have a definite charge parity.

Again, QED respects charge parity, and again, later on theories will be encountered, where this is not the case.

4.8 CPT

It is now a very deep result that quantum-field theories on Minkowski space-time have the following property: A combination of charge parity transformation, parity and time reversal, i. e. the application of the operator CPT (or any other order, as they commute) will always be a symmetry of the theory[6]. Especially, causality is deeply connected with this property, and one implies the other.

This implies that if one mirrors physics and exchanges particles by anti-particles, and runs everything backwards there will not be any difference observed in any physical process.

[6]There are some attempts, whether it is possible to formulate quantum-field theories which do not respect this, as experimentally there are only upper limits for any violation. However, so far none emerged with a convincing concept of how this really defines a reasonable field theory. It is, of course, possible that when going beyond the framework of quantum-field theory such possibilities may exist.

From this, it can also be shown that necessarily all integer spin particles respect Bose-Einstein symmetry, while all half-integer spin particles will respect Fermi-Dirac statistics. The connection is made using special relativity: On the one hand it implies that only certain spins may exist, and at the same time it implies that certain properties under the exchange of particles have to be respected. Connecting this leads to the aforementioned classification. Therefore, in the context of particle physics, integer-spin particles are synonymously called bosons and half-integer spin particles fermions. Since this is a particular property of three-(or higher) dimensional quantum-field theory endowed with special relativity on flat Minkowski space-time, which is the arena of the standard-model, this association is fixed in particle physics. One must be careful, if one moves to a different field.

As a consequence of this so-called CPT-theorem the behavior of a state or field under one of the symmetries is fixed once the other two are given. Conventionally therefore only the quantum numbers under P and C are recorded for a particle, which leads to the J^{PC} classification, where J is the total, orbital and intrinsic spin, angular momentum of a particle, and P and C are only denoted as $+$ or $-$. E. g. a photon has $J^{PC} = 1^{--}$. Of course, if any of the quantum numbers is not well-defined, e. g. the charge parity quantum number of particles with anti-particles of half-integer spin is not entirely trivial, it is omitted.

4.9 Symmetry breaking

Before continuing to further symmetries a brief intermission about breaking symmetries is necessary. Take a Lagrangian $\mathcal{L}_0$ with a symmetry at the classical level. There are now four possibilities how this symmetry may be affected.

The first is a somewhat obvious possibility, the explicit breaking. In this case an additional term is added to the Lagrangian, $\mathcal{L}_0 + \delta\mathcal{L}_b$, with δ a parameter, which does not obey the symmetry. Then, of course, already at the classical level the theory no longer has the symmetry. It can be shown that such an explicit breaking can only be performed for global symmetries. Any attempt to do this for a local symmetry, though classically possible, cannot be quantized consistently.

The relevance of explicit symmetry breaking comes about in two special cases. One is that of soft symmetry breaking. In this case the terms in $\mathcal{L}_b$ are such that they are only taking effect in a certain energy regime, but become negligible in another one. The most common case will be seen to be mass terms, which break symmetries of most particle physics theories at low energies, but become irrelevant at large energies. Therefore, if the theory is probed at sufficiently large energies, the symmetry appears to be effectively restored.

The second possibility is if δ is small. In this case the symmetry may not be exact, but relations due to the symmetry may still hold true approximately also in the full theory. In such a case it is said that the symmetry is only weakly broken. This concept plays an important role later for the strong nuclear interaction.

The next possibility is that a classical symmetry is no longer a symmetry of the quantized theory. Again, this only leads to a consistent quantum theory if the affected symmetry is a global one. Such an effect is also called an anomaly. In that case the symmetry is just not present at the quantum level. There are several examples to be encountered in particle physics. The simplest one is given by massless QED. The classical theory is invariant under a scale transformation, i. e. when rescaling all dimensionful quantities by a fixed number. This symmetry, the so-called dilation symmetry, is broken in the quantization. As a consequence, the theory has an intrinsic mass scale. This mass scale is e. g. the one encountered in the running of the coupling, see section 3.8. If the dilatation symmetry would be unbroken, the coupling would be independent of energy. Incidentally, it can be shown that this is deeply connected to the masslessness of the photon.

If a symmetry exists both classically and survives the quantization process, it may still be broken spontaneously. It can be proven that this cannot occur in usual particle physics theories for local symmetries, but only for global symmetries. This will manifest itself in the possible outcomes of the theory. E. g., if a symmetry predicts that two states should be the same, they will no longer be after spontaneous symmetry breaking. The spontaneous magnetization of a magnet below the Curie temperature is an often cited example for this phenomenon. Again, there is some fine-print in a full quantum-field theory[7], but the simpler picture will suffice for this lecture.

The fourth possibility is that the symmetry remains at the quantum level.

A more detailed treatment of what symmetry breaking really is will follow in section 5.7.

[7] E. g. the total magnetization of a magnet is always zero, because there is no preferred direction in space-time. Spontaneous magnetization manifests itself rather in the relative magnetization of two neighboring spins in the magnet. These are uncorrelated without spontaneous magnetization, but correlated in the case of spontaneous magnetization. Also, in the symmetry-broken case the system is unstable against external magnetic fields, and will even for an arbitrarily small external field develop a preferred magnetization direction.

4.10 Fermion number

An inspection of the QED Lagrangian (3.9) shows that there is another symmetry. If the fermion field is multiplied by a constant phase $\exp(i\alpha)$ and the anti-fermion with $\exp(-i\alpha)$, the Lagrangian remains unchanged. This is a continuous U(1) symmetry, and therefore a Noether current has to exist. Calculating it, its is found to be proportional to the electric current (4.4). However, it is not the electric current itself, but is actually the current of fermion number, which just in the particular case of all fermions having the same electric charge coincides with the former.

Therefore, the number of fermions, i. e. the difference of the number of fermions minus the number of anti-fermions, is conserved. This does not require that these fermions remain of the same type. E. g. a state with one electron or with one muon both have the same fermion number, one. Any process respects this property. In fact, within the standard model this number of fermions is always respected.

Note that there is no similar condition for photons, and they can be created and destroyed at will, provided other conditions like conservation of total angular momentum and four-momentum conservation are satisfied. This is not so because they are bosons, but because they are their own anti-particles. Later on bosons will be encountered, which carry charge, and there charge conservation will restrict this possibility. Still then no analogue in the form of a boson number conservation exist.

4.11 Flavor

QED has, however, not only a single fermion species, but three, so-called flavors. Multiplying only one of them with a phase still is a symmetry of the Lagrangian. Each of these three symmetries has its own conserved current. Therefore, the total numbers of electrons, muons, and tauons are conserved in QED separately, it constitutes a flavor quantum number. This flavor quantum number is, like electric charge or fermion number, an internal symmetry, and therefore flavor is an internal quantum number. QED is said to respect flavor symmetry, where the flavors are uniquely identified by their mass. Formally, this is a $\mathrm{U}(1)^3$ symmetry, i. e. a product of three independent U(1) groups.

If the three flavors would be mass-degenerate, this symmetry would be enlarged. Since all particles have the same mass and the same electric charge, they become indistinguishable, and they can also be exchanged. Thus, it is possible to perform rotations in the internal flavor space. If the leptons would be described by real fields, this would be just the usual orthogonal group of three-dimensional rotations, O(3). But because they are complex, this group is enlarged to SU(3). The total flavor symmetry of mass-degenerate QED is hence SU(3), or, when supplemented by the additional fermion number symmetry, SU(3)×U(1)∼U(3).

4.12 Chiral symmetry

The QED Lagrangian (3.9) exhibits one interesting additional symmetry if the masses of all the fermions are set to zero, besides the then manifest flavor symmetry. This additional symmetry emerges by the combination of a flavor or fermion number transformation and an axial transformation. Axial transformations are a special property of fermions, and there is no analogue for bosons of arbitrary spin. The combination of fermion number and the axial transformation is the simplest case. It is mediate by multiplying every fermion field by the matrix $\exp(i\alpha\gamma_5)$, where α is a real parameter, and $\gamma_5 = -i\gamma_0\gamma_1\gamma_2\gamma_3$ is a combination of the γ-matrices. This can be shown using the fact that γ_5 anti-commutes with all γ_μ. The anti-fermion field is transformed by the corresponding hermitian conjugated phase factor.

This phase symmetry adds an additional U(1) symmetry to the theory, which is called an axial symmetry $\mathrm{U}_A(1)$. In addition, like the generalization of the fermion number symmetry U(1) to the flavor symmetry $\mathrm{SU}(N_f)\times\mathrm{U}(1)$ for N_f flavors, it is possible to enlarge the axial symmetry to an axial flavor symmetry, called the chiral symmetry. This name stems from the fact that it turns out that it connects fermions with spin projections along and opposite to their momentum direction, i. e. of different helicities. Since these projections yield classically a left-handed and right-handed screw[8], the name chiral, Greek for handedness, is assigned. The total symmetry of the theory is therefore $\mathrm{SU}(N_f)\times\mathrm{SU}_A(N_f)\times\mathrm{U}(1)\times\mathrm{U}_A(1)$ for N_f flavors of leptons.

Of these symmetries, the axial symmetry is actually broken by an anomaly during quantization. Non-zero lepton masses break the chiral symmetry, and the non-degenerate lepton masses then finally break the flavor symmetry just to a flavor number symmetry, $\mathrm{U}(1)^{N_f}$. Hence, little is left from the classical symmetries of massless QED.

This symmetry is therefore not realized in nature, where these masses are non-zero and different. Furthermore, since the masses are large compared to the strength of electromagnetic interactions, the symmetry is so badly broken that it has almost no relevance for QED. This will change later when discussing the strong nuclear interaction.

[8]Note that the question of what is left-handed or right-handed depends upon whether you look at a screw from the top or the bottom. Since different conventions on how to look at a screw are in use, care should be taken.

4.13 Generators

A final remark on symmetries is the notion of generators, as it appears throughout the literature. It is a technically very useful observation that for a continuous symmetry, which changes $\phi \to \phi + \delta\phi$, the associated Noether charge Q plays a special role. It can be shown that

$$[\phi, Q]_{\pm} \sim \delta\phi,$$

where additional constants of proportionality may or may not appear, depending on conventions, and the $\pm$ indicates that this may be a commutator or anti-commutator, depending on whether the involved quantities are bosonic or fermionic, respectively. As a consequence, the charges Q are also referred to as the generators of the symmetry. In this sense, the electric charge creates the electromagnetic (gauge) symmetry.

Chapter 5

Strong interactions

It was very early on recognized that QED cannot be used to describe atomic nuclei, since QED can never sustain a bound state of the positive charge carriers which make up the nuclei, the so-called protons with their mass of roughly 938 MeV, and electrically neutral particles, the neutrons, with their just about 1.5 MeV larger2 mass. Since also gravity was not an option, there had to be another force at work, which is called the strong nuclear force. This became an ever more pressing issue, as it became experimentally clear that neither protons nor neutrons can be elementary particles, as they have a finite extension, about 1 fm, and behave in scattering experiments like having a substructure. To understand the details was, however, a rather complicated challenge.

The formulation of the strong interactions in its currently final form of quantum chromo dynamics (QCD) is, in a sense, the latest addition to the standard model. Because of the phenomenon of confinement, discussed below in section 5.8, it was much longer formulated than experimentally established. The reason behind is that the strong nuclear force is, as it name suggests, strong. It is therefore only in very limited, though relevant, circumstances possible to use perturbation theory to calculate anything in QCD. This made theoretical progress slow. The mainstay of QCD calculations today are in the form of numerical simulations, which for many problems in QCD physics has become the method of choice. Still, since computational power is limited, though such simulations being among the top ten contender for computation time in the world, it is not yet possible to calculate complex objects like even a helium nucleus reliably. Hence, for many purposes, many of the more fundamental features of the theory remain an area of active research today. In this context, QCD and theories which are (slightly) modified versions of QCD also serve as role models for generic strongly interacting theories in particle physics.

5.1 Nuclei and the nuclear force

As noted, all atomic nuclei consists out of differing numbers of protons and neutrons, the former having a positive electric charge of the same size as the electron and the latter no electric charge at all. The simplest atomic nucleus is a single proton. However, single neutrons are not stable, but decay in a few minutes, and are therefore only appearing bound in nuclei.

This has been found, once more, by scattering experiments. When scattering with an electromagnetic probe, e. g. an electron, the characteristic quantity is the total momentum transfer $-q^2 = Q^2$ between the electron and the nuclei, usually transmitted by a photon. Defining furthermore the fraction of momentum

$$x = \frac{Q^2}{2pq},$$

where p is the momentum of the target. In the rest frame of the target the denominator becomes Mq_0, where the transferred energy q_0 is also often called the energy transfer ν. The variable x is called the Bjorken variable, or Bjorken-x, for short. Since in a two-body collision at fixed energy Q^2 is entirely characterized by a scattering angle, a description in terms of x is as well suited.

In the case of elastic scattering, x becomes one. Hence, the cross-section, or, alternatively, the structure function, as a function of x will be just a single peak at 1. Increasing the energy transfer, x becomes smaller than one. Then, additional peaks will appear, which correspond to excited states of the nuclei. Such excited states already show the presence of an internal structure, together with the finite extension of the nuclei. In addition, around $x \approx 1/A$, with A the number of nucleons in the nuclei, there is a broad peak. This broad peak is due to the elastic scattering of the nucleons in the nuclei. The broadness comes from the Fermi motion, i. e. from the fact that a nucleon confined to a nuclei has, due to the uncertainty relation, at least a momentum of roughly $1/R$, where R is the nuclei radius. With $R - 1 - 10$ fm, these are energies of order 100 MeV, and therefore not negligible on nuclear energy scales.

Because of electromagnetic repulsion it became quickly clear that the nuclear force must be very much stronger than QED to create quite compact nuclei, not much larger than the protons and neutrons, with more than one proton. It must also act on different charges, as the neutron is electrically neutral. One additional observation was made, when investigating nuclei and the nuclear force. It was not only because of its strength very different from QED and gravity, but it was also not of an infinite range. Though being much stronger than electrodynamics, it dropped almost to zero within a very short range of a few Fermi.

How can such a short-range force emerge? One suggestion comes from the relativistic description of bosons. Take the Klein-Gordon equation (1.3) for a scalar boson of mass m,

$$(\partial^2 + m^2)\psi = 0, \tag{5.1}$$

with the wave-function ψ as the solution. Investigating a static situation yields

$$\psi(\vec{r}) = \frac{g}{4\pi|\vec{r}|}e^{-m|\vec{r}|}, \tag{5.2}$$

where g is an integration constant. Thus, the wave-function decays exponential towards large distances, where the characteristic distance of propagation is given by the inverse mass of the particle. Hence, a short-range interaction can arise if the exchanged particles are massive. This lead to the prediction of an exchange boson for the nuclear force with a mass of the order of roughly hundred MeV, given the range of the nuclear force of about a few fm. Such a potential is called a Yukawa potential.

The interpretation of the integration constant g becomes clear when the zero-mass limit is considered. In this case the wave-function becomes Coulomb-like, with g corresponding to the fine-structure constant. Thus, also in the massive case g can be interpreted as a coupling constant, which characterizes the strength of the strong nuclear force. It turns out that characteristic values for it are about two orders of magnitude larger than for the electromagnetic force.

Before taking this as more than a hand-waving motivation to look for a massive particle, one should add a word of caution with respect to the Klein-Gordon equation. It is, strictly speaking, only valid if the energy scales involved are small compared to the mass. Furthermore, it does not lead to a stable vacuum state, and therefore it can only describe physics relative to an (undetermined) vacuum. Both sicknesses are only cured when moving to quantum field theory, though this is beyond the scope of this lecture.

Nonetheless, the Yukawa potential yields the right idea. Instead of having massless photons as force particles, massive particles must mediate the strong nuclear force. They were indeed found in the form of the mesons.

5.2 Mesons

While the protons and neutrons are fermions with spin 1/2, the force carrier of the nuclear force were identified to be actually bosons. The lightest of them are the pions with quantum numbers $J^P = 0^-$, i. e. they are pseudoscalars. They come as a neutral one, π^0, and two oppositely charged ones, $\pi^\pm$. The range of the nuclear force is about 1 fm, which indicates that the mass of the force carrier, according to the Yukawa potential (5.2), should have a mass around 100-200 MeV. Indeed, the pions are found to have masses of 135.0 and 139.6 MeV for the uncharged and charged ones, respectively, and are thus much lighter than either protons or neutrons. These pions are not stable, but decay either dominantly electromagnetically into photons for the neutral one or like the neutron for the charged ones. Their life-time is of the order of 10^{-8} seconds and 10^{-17} seconds for the charged and uncharged ones, respectively. Therefore the charged ones live long enough to be directly detectable in experiment.

One of the surprises is that the neutral one decays into two photons, as usually photons are expected to couple only to electromagnetically charged objects. While this can be thought of as a neutral pion virtually splitting into two charged pions, and then annihilation under emission of photons, this is somewhat awkward. A more elegant resolution of this will be given in the quark model below in section 5.4.

With these pions it was possible to describe the overall properties of nucleons, especially long-range properties. At shorter range and for finer details it turned out that a description only with pions as force carriers was impossible. This was resolved by the introduction, and also observation, of further mesons. Especially the vector meson ρ with a mass of 770 MeV, spin one, and a very short life-time of roughly 10^{-24} seconds and the vector meson ω with a mass of about 780 MeV, but with a 20 times longer life-time than the ρ, play an important role. This larger number of mesons is also at the core of apparent three-body forces observed in nuclear interactions, which are, e. g., necessary to describe deuterium adequately. In fact, many more mesons have been discovered, and some more will appear later.

Describing how these various mesons create the strong nuclear force is in detail very complicated, but it in principle can be systematically performed, e. g. in the form of the so-called chiral perturbation theory. This will lead too far astray from particle physics itself, and will therefore not be detailed here. What is, however, remarkable is that out of nowhere appear several different mesons, all contributing to the nuclear force, and actually all of them also affected by the nuclear force. Such a diversity of force carriers is distinctively different from the case of QED, where only the photon appears.

5.3 Nucleons, isospin, and baryons

In the endeavor to find the carriers of the nuclear force, several other observations have been made. The first is that most nuclear reactions show an additional approximate symmetry, the isospin symmetry. This symmetry is manifest in the almost degenerate masses of the proton and the neutron. Both particles can therefore be considered to be a doublet of this new symmetry, which is similar to the flavor symmetry of QED, just much less broken because of the very similar masses. It is furthermore found that also the three pions fit into this scheme.

That once two particles and once three particles appear can be understood in the following way, very similar to the case of angular momentum or spin in quantum mechanics. There the rotational symmetry existed, and states could have any total spin l, and then there existed a degeneracy of the states into $2(l+1)$ states with different third components. Similar for the isospin, a state can have a particular isospin value I, and then there are different charge states. In fact, isospin can be related to an SU(2) group once more, just like spin in quantum mechanics, hence the name. The proton and neutron then belong to the case of isospin I being 1/2, and thus there are two states with different third components of the isospin I_3, which are essentially charge states. Such a situation is called a doublet, and therefore proton and neutron are subsumed under the name of nucleons. Since the symmetry is only approximate, the different charge states are not degenerate. This is similar to the Zeemann effect in quantum mechanics, only that the mass now takes the role of the magnetic field.

The pions then belong to a state with $I = 1$, and thus three states. This is called a triplet, with approximately the same mass. Generically, such collections are called multiplets. In group theory, they correspond to different representations.

The natural question is then, whether there are higher representation, e. g. $I = 3/2$, with four states. Based on the fact that isospin seems to be related to electric charge, since the different states of an isospin multiplet all have different charge, and that half-integer or integer isospin corresponds to fermions or bosons, the properties of such a quadruplet should be predictable. In fact, the relation

$$Q = I_3 + S \tag{5.3}$$

seems to hold so far, where I_3 is the charge of the state, and S its spin[1]. For the proton and neutron, the assignments $I_3 = 1/2$ and $I_3 = -1/2$, respectively, yield the correct result. For the three pions, the I_3 assignments of -1, 0, and 1 reproduce the correct electric charges for π^-, π^0, and π^+.

However, this fails for the ρ meson, which has spin one, and is an isospin singlet, but in contrast to the rule (5.3) is uncharged. This can be remedied by replacing the spin S by $B/2$, where B is the so-called baryon number,

$$Q = I_3 + \frac{B}{2} \tag{5.4}$$

[1]The spin will already in the next step be replaced by a different quantum number.

which is one for the nucleons and zero for the mesons. This is so far a phenomenological identification, but will become quite relevant in section 5.4. Of course, the anti-particle of the nucleons, the anti-proton and anti-neutron, carry negative baryon number.

According to this rule, it is possible to attempt to construct a quadruplet, having four states with $I_3 = -3/2, -1/2, 1/2, 3/2$. To get integer charges, it must then have a baryon number, like the nucleons. These particles should therefore have electric charge -1, 0, $+1$, and $+2$, and corresponding anti-particles. These particles have been observed experimentally, and again the different states have almost the same mass. They are called Δ, have masses of about 1232 MeV, and are fermions, as are the nucleons. However, their spin is 3/2. Since both nucleons and baryons carry baryon number, there are called commonly baryons, to distinguish them from the mesons. In fact, they are not the only baryons, and many more have been found experimentally.

Together, mesons and baryons are denoted by hadrons, and are identified as those particles directly affected by the strong nuclear force.

5.4 The quark model

The number of baryons and mesons found by now numbers several hundreds. Already decades ago, when only a few dozens were known, it appeared unlikely that all of them should be elementary. This was very quickly confirmed by experimental results which showed that the proton had a finite size of about 1 fm, and Rutherford-like experiments found that there are scattering centers inside the proton, which appeared point-like. In fact, similar to the case of the nuclei, the experiments showed, to leading order, that the number of constituents in baryons is three, and in mesons it is two. However, beyond leading order, things were quite different. In the case of nuclei, as long as the energy transfer Q^2 is not large enough to disintegrate the nuclei, the number of constituents is essentially constant, and the elastic peak of the sub-structure scattering is essentially always at $1/A$, independent of Q^2. This turned out not to be the case for the nucleons. Depending on Q^2, the peak started to shift, and an increase is found towards small x, even though some peak structure remains at a third and a half, for baryons and mesons, respectively. This would indicate the presence of many more constituents, called partons, if the nucleon is probed at large energies. This kind of so-called scaling violation cannot be obtained from an interaction like (5.2), requiring a new type of interaction.

In addition, in contrast to the nuclei, these constituents were not almost isolated in essentially free space, but very tightly packed. Furthermore, while the neutron is electrically neutral, it was found to have a magnetic dipole moment, a feature beforehand believed to be only existing if there is an electrically charged substructure present.

This evidence together suggested that the elementary particle zoo could possibly be obtained from simpler constituents and put into a scheme like the periodic table of chemical elements, which originates from just three different particles.

Playing around with quantum numbers showed a number of regular features of the hadrons. This gave rise to the quark model, where in the beginning two quarks were needed to explain the regularities observed, the up quark and down quark, abbreviated by u and d, as well as their anti-particles. Since both the bosonic mesons and fermionic hadrons must be constructed from them, it requires them to be fermions themselves. Since all of the hadrons have an extension, none of them can be identified with a single quark, just like the periodic table does not contain a single proton, neutron or electron. However, in contrast to the latter no free quarks are observed in nature, a phenomenon to be discussed below in section 5.8.

The simplest possibility to construct then a hadron would be from two quarks. This must be a boson, as two times a half-integer spin can only be coupled to an integer spin, and therefore a meson. Since no free quarks are seen, the nucleons must contain at least three quarks to get a half-integer spin. These considerations turn out to be correct. However, they lead to the conclusion that quarks cannot have integer electric charges. This is most easily seen by looking at the nucleons.

Scattering experiments identified that the nucleons have no uniform sub-structure, but have a two-one structure, that is two quarks of one type and one quark of the other type. These two types were denoted up and down type, in correspondence to their isospin values. Since it is found that the down quark is heavier than the up quark, the heavier one, i. e. the neutron, should have two down quarks. This yields a composition of uud for the proton and udd for the neutron. The only solution for the observed electric charges of the proton and the neutron are then an assignment of $2/3$ of the (absolute value of the) electron charge for the up quark, and $-1/3$ of the (absolute value of the) electron charge for the down quark. This consistently yields the required positive and neutral proton and neutron charges. This also explains the magnetic dipole moment of the neutron. At the same time, the baryon number of quarks must be $1/3$ for up quark and down quark. This implies also that the isospin of the up quark is $+1/2$ and that of the down quark is $-1/2$.

The pions are then constructed as a combination of a quark and an anti-quark, $u\bar{d}$ for the π^+, $\bar{u}d$ for the π^-, and a mixture of $\bar{u}u$ and $\bar{d}d$ for the π^0, i. e. the state of the π^0 is

$$\left|\pi^0\right\rangle = \cos\alpha\,|\bar{u}u\rangle + \sin\alpha\,\left|\bar{d}d\right\rangle,$$

where α is a mixing angle. This mixing angle can be experimentally or theoretically determined. Experimentally, this is possible by using the different decay patterns for different mixing angles. Theoretically, by calculating the left-hand side and the right-hand side in different bases. However, in practice both possibilities are highly challenging. An assignment of two quarks instead of a quark and anti-quark is not possible, as this cannot give the required baryon number of zero.

Particles like the ρ meson are then also combinations of a quark and anti-quark, but where the quarks have relative orbital angular momentum, creating their total spin of one.

The fact that hadrons cannot be elementary has already been established much earlier, already in 1933. If protons and neutrons would be elementary particles then they should have, up to small electromagnetic corrections, a magnetic moment similar to that of the electron, just modified by their mass,

$$\vec{\mu}_N = \frac{q}{2m_N}\vec{\sigma},$$

where m_N is the respective mass, q the charge, and $\vec{\sigma}$ the normalized direction of the spin. Thus, the neutron should have zero magnetic moment and for the proton of $\mu_N \approx q/2m_N$. The experimental result is very different. The proton has $2.79\mu_N$ and the neutron $-1.91\mu_N$.

This can be understood in the quark model. The resolution is that the total magnetic moment is combined from the magnetic moments of the constituent quarks. Assuming for now their masses to be about a third of the nucleon mass, their respective magnetic moments are $2/3 \times 1/1/3\mu_N \approx 2\mu_N$ for the up quark and $-1/3 \times 1/1/3\mu_N \approx -\mu_N$ for the down quark. However, the total spin is given by the coupling of the quark spins. Since the total spin of the nucleons is 1/2, two of the quarks have to be coupled to a singlet.

For the proton, this singlet can be either two ups or an up and a down. In the first case, the two ups can be distinguished, in the latter not. The result is then

$$\frac{2}{3}(2\mu_u - \mu_d) + \frac{1}{3}\mu_d = \left(\frac{2}{3}(4+1) - \frac{1}{3}\right)\mu_N = 3,$$

where the factor 1/3 is from the normalization of the wave function. For the neutron the same applies, except that ups and downs are exchanged,

$$\frac{2}{3}(2\mu_d - \mu_u) + \frac{1}{3}\mu_u = \left(\frac{2}{3}(-2-2) + \frac{2}{3}\right)\mu_N = -2.$$

These results are already pretty close to the experimental ones, there is just a 10% deviation left.

However, not all particles fit into this simple model. The Δ turns out to pose a serious challenge.

5.5 Color and gluons

At first glance, the Δ appears simple enough. The double-positive state Δ^{++} is just three up quarks, and with decreasing charge always one up quark is replaced by one down quark, until reaching the Δ^- with three down quarks. To obtain the observed 3/2 spin requires to align the spin of all three quarks. Of course, it could be possible that there would be a relative orbital angular momentum, but experimentally this is not found. In fact, there exists an excited version of the Δ with such an orbital angular momentum and total angular momentum of 5/2, which is also experimentally confirmed.

And this is where the problem enters. Since the Δ is a fermion, its wave-function must be totally antisymmetric. Since the spins are aligned and all three quarks are of the same type in the ground-state, no wave-function can be constructed which is anti-symmetric. Thus the existence of the Δ appears to violate the Pauli principle at first sight. But this is not so. Originally introduced to resolve that problem, and later experimentally verified, another conserved quantum number is attached to quarks: Color. The wave-function can then be anti-symmetric in this new quantum number, saving the Pauli principle and the quark model at the same time.

Since this new quantum number of the quarks is not observed for the Δ, or any other hadron, the hadrons must all be neutral with respect to this new quantum numbers. For the mesons, consisting of a particle and an anti-particle, this is simple enough, as just both have to have the same charge. This is not the case for baryons. Assigning just positive or negative charges, like the electrical charge, it is not possible to construct neutral states out of three particles. Attempts to do so with fractional charges also do not succeed in the attempt to make the proton and neutron color-neutral simultaneously. It is therefore necessary to depart from the simple structure of the electromagnetic charge.

As a consequence, it is assumed that there are three different charges, suggestively called red, green (or sometimes, especially in older literature, yellow), and blue. It is furthermore assumed that not only a color and the corresponding anti-color is neutral, but also a set of each of the colors is neutral. Then there are three quarks for each flavor: red, green, and blue up quarks, and red, green, and blue down quarks, totaling six quarks. A color-neutral baryon is containing a quark of each color, e. g. a proton contains a red and a blue up quark, and a green down quark. In fact, since the total charge of a proton is zero, it is a mixture of any possible combination of color assignments to each of the three quarks, which are consistent with neutrality and the Pauli principle. Similar, the Δ^{++} consists of a red up quark, a green up quark, and a blue up quark.

This construction is rather strange at first sight, but it can be formulated in a mathematically well-defined way. This will be done below in section 5.16, after collecting the other ingredients of the strong interactions.

One other important ingredient, now that there is a new charge, is what mediates the force between the charges. In electromagnetism it was the massless photon. It is therefore reasonable to assume that there is also a mediator of the force between color charges. These were indeed found, and named gluons. As the photon these are massless[2] bosons with spin one. However, they differ from photons in a very important property. While photons are only mediating the electromagnetic force, they are not themselves affected by it, since they carry no electric charge. But gluons carry color charge. In fact there are 8 different charges[3] carried by gluons, and none of these eight are either the quark charges, nor is there any simple relation to the quark charges. Especially, it is impossible to add a single quark charge with any combination of the gluon charges to obtain a neutral object. To achieve this, at least two quarks have to be added to one or more gluons.

Nonetheless, the idea of gluons has been experimentally verified, and they have been identified as the carrier of the strong interaction, binding quarks into color-neutral hadrons. The exchange of mesons to bind nucleons into nuclei can be viewed as a high-order effect of the gluon interaction. This is similar to Van-der-Waals force, though the details are different, as here not a color dipole moment enters, and the details are not yet fully resolved. Still, the interaction of nucleons in a nuclei can be traced back to the gluons.

Hence, the combination of quarks, gluons, and colors can explain the structure of all known hadrons, similar to the periodic table. Unfortunately, the strong force binding quarks by gluon exchange is not accessible using perturbation theory, at least when it comes to describing hadrons. Its treatment is therefore highly non-trivial. Because of the color, this underlying theory of hadrons is called chromodynamics, its quantum version quantum chromodynamics, or QCD for brief.

That this construction makes sense can also be read off from experiment. Consider the ratio

$$\frac{\sigma_{e^+e^-\to\text{Hadrons}}}{\sigma_{e^+e^-\to\mu^+\mu^-}}$$

as a function of the center-of-mass energy s. To leading order, this will occur because the lepton pair is annihilated into a photon, which then splits either into two quarks or two leptons. Thus, the difference comes from the number of quarks. If there are three quark

[2]Again, the notion of mass is murky here, and the precise details are only becoming clear in a full non-perturbative treatment of the corresponding quantum field theory. However, for most purposes the notion of massless is sufficient.

[3]One can think of four of them as particles and four as anti-particles, though this is not really how it works, see section 5.16.

colors, this ratio will behave for not too large ($\sqrt{s} \lesssim 3.2$ GeV) energies as

$$3\left(\frac{1}{9}+\frac{4}{9}+\frac{1}{9}\right)=2.$$

Without color, this ratio, also known as R ratio, would be smaller by a factor of three. This is experimentally well confirmed, though resonance structures due to exchange of other particles obscure the picture in some ranges of s from the simple underlying mechanism.

5.6 Chiral symmetry breaking

No word has yet been said about how the masses of the hadrons relate to the masses of the quarks. The mass of the proton is known very precisely to be 938.3 MeV, and the proton to be 939.6 MeV. This implies that the mass difference between up and down quarks must be tiny. The Δ is somewhat heavier, about 1230 MeV. This can be understood as an excited state, and it is therefore heavier. Most ground state mesons have a mass of about 600 MeV or more. All this suggest a mass of about 300 MeV for the up quark and down quark, with very little difference[4]. But two mysteries appear. One is that the pions are very light, just about 140 MeV. The second is that any attempt to directly measure the quark masses yield consistently a mass of about 2.3(7) MeV for the up quark, and 4.8(5) MeV for the down quark. Though the difference is consistent, the absolute values are much smaller than the suggested 300 MeV from the nucleon properties.

The resolution of this puzzle is found in a dynamical effect of the strong interactions. To understand it, recall that massless free fermions have a chiral symmetry, see section 4.12. I. e. they can be multiplied by a phase factor $\exp(i\alpha\gamma_5\tau^a)$, where the τ^a are the Pauli matrices, if the two flavors of quarks are arranged in a vector. This symmetry is, at first sight, approximately valid for up and down quarks as their intrinsic masses (often called current mass) is very small compared to other hadronic scales, like the hadron masses. Thus, naively it is expected to hold. This chiral symmetry has a number of particular consequences. E. g. it implies that bound states of opposite parity, but otherwise identical content, should have the same mass. But the mass splitting between such bound states for hadrons is large. E. g., the parity partner of the nucleon is called the $N(1535)$ and has, as its name suggests a mass of 1535 MeV, about 50% heavier than the nucleons. This is much larger than expected due to the explicit breaking of the chiral symmetry because of the small current masses.

[4]Though this difference is crucial for making the proton lighter and thus (more) stable (than the neutron), and therefore chemistry possible.

Thus, chiral symmetry must be much stronger broken than just from the current quark masses. The strong interactions must either spontaneously or explicitly break chiral symmetry. Experiments show that at sufficiently high energies, many times the hadronic scales, the quarks again behave as if they have only their current masses. Thus, the breaking must be spontaneous. This will be confirmed when writing down the field theory of QCD in section 5.16. Hence, the strong interaction break chiral symmetry spontaneously.

This proceeds in the following (hand-waving) way. The gluon interaction creates a strong binding of the quarks, which creates a so-called vacuum condensate[5] $\langle\bar{u}u\rangle + \langle\bar{d}d\rangle$. Since the condensate is made from the quarks, any other quark will interact with it. Since the strong force is attractive, otherwise protons, neutrons, or nuclei would not exist, this will slow down any movement of a quark. Thus, quarks gain in this process inertial mass[6], and thus effectively a larger mass. This additional mass is the same for up quarks and down quarks, and approximately 300 MeV. This explains how the heavier mesons and the baryons gain a mass of this size. It does not yet explain two other facts: The lightness of the pions nor why this is no longer the case at high energies. These two questions will be answered in the next two sections.

Before this, some remarks should be added. It is easy to wonder how the universe should be filled with such a condensate, but no effect appears to be visible. Here, one notes how everyday experience can be deceiving. Of course, the effect is visible, since without it, all nucleons would be very much lighter, and so would be we, since almost all our mass is due to nucleons. Thus, whenever we feel our mass, we feel the consequences of this condensate. In fact, this condensate is responsible for essentially 99% of the mass of everything we can see in the universe, i. e. suns, planets, asteroids, and gas, the so-called luminous mass of the universe.

5.7 The Goldstone theorem

The lightness of the pions is a very generic feature of particle physics theories with spontaneous symmetry breaking. It is formulated in Goldstone's theorem. To lay out this theorem, it is helpful to investigate a very simplified model.

Take two scalar particles[7], which can interact with themselves. Such particles can either be described by two scalar fields ϕ_1 and ϕ_2 or with a single complex field $\phi = \phi_1 + i\phi_2$.

[5] The omission of the state generally denotes that the expectation value is taken with respect to the vacuum.

[6] That this also acts as a gravitational mass is not at all clear. Only the equivalence principle implements this as a law of nature, an axiom.

[7] With only one scalar field it is not possible to construct a simple example with a continuous symmetry.

Such a theory has a classical Lagrangian of

$$\mathcal{L} = \partial_\mu \phi^\dagger \partial^\mu \phi - U(\phi^\dagger \phi), \tag{5.5}$$

where U describes the potential for the field, which can only come from interactions between the two fields. The simplest (and in known particle physics so far exclusively appearing) possibility to obtain a symmetry for this theory is that the potential depends only on the product $\phi^\dagger\phi$.

To obtain the simplest example for a symmetry, take the potential

$$U(\phi^\dagger\phi) = \frac{\mu^2 v^2}{4} - \frac{\mu^2}{2}\phi^\dagger\phi + \frac{\mu^2}{4v^2}(\phi^\dagger\phi)^2.$$

The pre-factors, i. e. coupling constants, as well as the irrelevant constant term have been chosen judiciously such that the result will be looking simple. This potential, as well as the kinetic term, is invariant under the phase rotation $\phi \to \exp(i\alpha)\phi$. Therefore, this theory has a global (U(1)) phase symmetry. It is a bit odd theory, as the quadratic term is usually associated with a mass. This can be seen when determining the classical equation of motion, which reads

$$0 = \partial_\mu \frac{\partial \mathcal{L}}{\partial \partial_\mu \phi^\dagger} - \frac{\partial \mathcal{L}}{\partial \phi^\dagger} = \partial^2\phi - \frac{\mu^2}{2}\phi + \frac{\mu^2}{2v^2}\phi^2\phi^\dagger. \tag{5.6}$$

The equation of motion for the field $\phi^\dagger$ is, up to conjugation, identical. Ignoring the interaction term, this is just again the Klein-Gordon equation (5.1), which describes the movement of a free scalar particle, if the sign of the second term would be positive. In this case, it looks like the particle has an imaginary mass, what would be called a tachyon. This is, however, not a problem. The sign of the quartic term is positive. Therefore, the energy is bounded from below classically, and the theory remains stable. It is therefore just an odd term in the potential energy.

To proceed, an interesting question is what the classical lowest energy state is. Since the kinetic term is positive, any spatial or temporal variation would increase the total energy. Hence, the state of lowest energy is necessarily a field ϕ_0 constant throughout space and time, which minimizes the potential (5.6). This constant is found to be $\phi_0 = ve^{i\theta}$, where θ is an arbitrary phase, i. e. all values of θ are a minimum. Thus, the solution manifold is highly degenerate. This is a consequence of the symmetry: Any change in θ can be offset by a symmetry transformation, without changing the physics.

One can proceed by specifying θ further. However, any choice is physically indistinct, and therefore arbitrary. For further calculations keeping θ manifest is awkward, and therefore in the following the explicit choice $\theta = 0$ is made. Since the choice is arbitrary, the symmetry is not violated. However, the symmetry is no longer manifest either. One therefore speaks of hiding the symmetry, or a hidden symmetry. However, it is hence actually an abuse of language, though a common one, to call it a broken symmetry.

To make the situation more transparent, the next step is to shift the field ϕ by its vacuum expectation value, i. e. replace

$$\phi(x) \to v + \eta(x) + i\xi(x), \tag{5.7}$$

making now the complex structure manifest with the two real fields η and ξ. In this way, fluctuations of the fields around the classical vacuum can be studied. Inserting this into the Lagrangian (5.5) yields

$$\mathcal{L} = \partial_\mu\eta\partial^\mu\eta + \partial_\mu\xi\partial^\mu\xi - \mu^2\eta^2 + \frac{\mu^2}{v}\eta^3 + \frac{\mu^2}{v}\eta\xi^2 + \frac{\mu^2}{2v^2}\eta^2\xi + \frac{\mu^2}{4v^2}\eta^4 + \frac{\mu^2}{4v^2}\xi^4. \tag{5.8}$$

This Lagrangian shows now a number of very interesting features, which are very generic.

The first is that the two fields η and ξ behave differently. While there is a mass term, now with the correct sign, for η, giving it a mass of $\sqrt{2}\mu$, there is no mass for ξ. Pictorially, one can think of η as excitations which describe fluctuations out of the minimum, while ξ, which is orthogonal to the direction of the chosen vacuum, moves between the different minima of the potential. Since the vacua all have the same energy, this does not cost any energy, and therefore the mode is massless. This is a generic feature of such situations, and is known as Goldstone's theorem. In a nutshell, it is the statement that there are as many massless particles as there are directions in which the minima are equivalent[8].

The second is that there are now many different interactions between the fields η and ξ. However their couplings, i. e. their pre-factors, are not all different, but completely determined by the original parameters. The reason is that the symmetry is just hidden. To ensure that any symmetry transformation is still valid requires that the various interactions cannot have arbitrary pre-factors, because otherwise it would no longer be invariant under the symmetry transformation

$$v + \eta(x) + i\xi(x) \to e^{i\alpha}(v + \eta(x) + i\xi(x)), \tag{5.9}$$

where it should be noted that the vacuum solution v is also transformed accordingly. For that to work out, it is necessary to keep track when changing from (5.5) to (5.8), which occurrence of v stem from the original coupling constants in (5.5), and which from the shift (5.7), since only the latter are affected.

[8]The precise formulation is that there are as many massless particles as there are generators of the symmetry group minus the number of generators of symmetry transformations after all possible remaining choices have been made, here this is one minus zero, and thus one massless particle.

If at any point a term is added to the Lagrangian, which violates the symmetry, the symmetry becomes explicitly broken. The most obvious way is to add a mass term for the ξ field to the Lagrangian (5.8). Then, the Lagrangian is no longer invariant under the symmetry transformation (5.9). Of course, this can be translated back into the original Lagrangian (5.5), where it takes the form of an additional quadratic term in the potential (5.6) of type $(\Im\phi)^2$. The effect is essentially that the potential is tilted, and the vacuum state has now a unique solution v, which has no longer an invariance. This gives also a physical explanation for the mass: Since there are no degenerate vacuum solutions anymore, any movement increases the energy.

The situation for the pions is the same, but just somewhat more complicated, since there are three pions. This can be described by an effective model, the so-called linear σ-model. In it, not only the pions are described, but an additional field, the so-called σ meson, which is a mixture of $u\bar{u}$ and $d\bar{d}$ in the quark model, a flavor-less (isosinglet) scalar. These are described by a real vector $\phi = (\pi^+, \pi^0, \pi^-, \sigma)^T$ inside the Lagrangian

$$\mathcal{L} = \frac{1}{2}\partial_\mu\phi_i\partial^\mu\phi_i + \frac{\mu^2}{2}\phi^2 - \lambda(\phi^2)^2. \tag{5.10}$$

This Lagrangian is invariant under O(4), i. e. under four-dimensional rotations of the vector ϕ by $\phi \to A\phi$ with A a four-dimensional rotation matrix, including mirroring $\phi \to -\phi$.

The first and second derivative of the potential term is

$$\begin{aligned} \partial_i U &= \phi_i(\mu^2 - 4\lambda\phi^2) \\ \partial_i\partial_j U &= -8\lambda\phi_i\phi_j + \delta_{ij}(\mu^2 - 4\lambda\phi^2) \end{aligned} \tag{5.11}$$

This yields extrema at $\phi_i = 0$, a maximum, and at $\phi^2 = \mu^2/(4\lambda)$, a minimum. Just as before, the Lagrangian can now be rewritten by separating the pions in a three-dimensional vector $\Pi = (\pi^+, \pi^0, \pi^-)^T$ and the single σ meson. This reduces the symmetry, yielding an O(3), i. e. three-dimensional rotation, of the pion vector. The Lagrangian is, up to constant terms, given by

$$\mathcal{L} = \frac{1}{2}\partial_\mu\pi^i\partial^\mu\pi^i + \frac{1}{2}\partial_\mu\sigma\partial^\mu\sigma - \mu^2\sigma^2 - \lambda(\pi^2)^2 - 2\mu\sqrt{\lambda}\pi^2\sigma - 2\lambda\pi^2\sigma^2 - 2\mu\sqrt{\lambda}\sigma^3 - \lambda\sigma^4.$$

Hence the pions are indeed Goldstone bosons, while the σ is massive with mass $\sqrt{2}\mu 2$. That the original O(4) symmetry is lost can be seen by the fact that such a rotation will mix the (zero) mass term of the pion and the mass term of the σ, and is therefore not a symmetry anymore.

To get that the pions have a non-vanishing mass, add to (5.10) a term $h\sigma^2$. This explicitly breaks the original O(4) symmetry directly to O(3). The additional term only modifies the derivative with respect to σ. The new minimum is then

$$\sigma_0 = \pm\sqrt{\frac{\mu^2 - 2h}{4\lambda}} = v.$$

This implies that $2h$ must be smaller than μ^2. The corresponding effective Lagrangian is

$$\begin{aligned}\mathcal{L} &= \frac{1}{2}\partial_\mu\pi^i\partial^\mu\pi^i + \frac{1}{2}\partial_\mu\sigma\partial^\mu\sigma - h\pi^2 - (2h+\mu^2)\sigma^2 \\ &- 2\sqrt{\lambda}\sqrt{2h+\mu^2}\pi^2\sigma - \lambda(\pi^2+\sigma^2)^2\end{aligned}$$

This implies that the pions have a mass $\sqrt{2}h$ and the σ of $\sqrt{4h+2\mu^2}$, which is thus heavier. Since h is arbitrary, in full QCD this term stems from the quark masses, and the pion can be massive, but arbitrarily light.

This illustrates how pions gain their small mass. The breaking of chiral symmetry would lead to a number of Goldstone bosons. In case of two quark flavors[9], there will be three Goldstone bosons, which are the three pions. These would be massless, if the quarks would be massless. However, because of the small current mass of the quarks, the symmetry is not exact, but rather explicitly broken. This gives the mass to the pions. That the masses of the pions are still large compared to the current masses of the quarks is a dynamical effect. Approximately, the pion masses scale quadratically with the current masses of the quarks, the so-called Gell-Mann-Oaks-Renner relation, but the pre-factor is large.

Before continuing on, a few words about subtleties and semantics must be said. When going to the quantum theory, quantum effects do what they always do, and they will mix all the degenerate vacuum states, and no vacuum will be preferred. Therefore, the vacuum state will exhibit a perfect symmetry, in contrast to the classical case. But the quantum system also carries the seed of the classical physics within, as it is metastable. If any arbitrarily small external perturbation, e. g. an infinitesimal mass for the ξ/σ from some other physics process, arises, it will immediately have a unique vacuum state, in which the symmetry is no longer realized. Hence, though strictly speaking the system without external influence is perfectly symmetric, the presence of this metastability has lead to the expression that the symmetry is nonetheless spontaneously broken.

In fact, even though the symmetry is exact, a full non-perturbative calculation shows that the system has both an ordinary massive and a massless excitation, and hence the most pertinent feature of the Goldstone theorem are realized even with the symmetry present. Of course, in a perturbative calculation this will not show. Since perturbation theory only permits very small deviations from the vacuum state, the particles will still

[9]It can be shown that the number of colors does not matter.

appear to be tachyons, as the relevant Lagrangian is (5.5). To cure this problem, one can introduce a weak external perturbation to the theory, which prefers a single vacuum, perform perturbation theory around this vacuum, i. e. using the Lagrangian (5.8) instead, and remove the external perturbation at the end[10]. Then, also in perturbation theory the system exhibits a massive and a massless particle. Especially because of this trick, which is extremely useful in the standard model, the exact notion of hidden symmetry is nowadays very rarely used, and almost always the situation will be denoted by spontaneous symmetry breaking.

5.8 Confinement and asymptotic freedom

This explains the light mass of the pion. It does not explain why the quarks appear to become lighter at high energies. The reason is that the strong interaction works very differently than electromagnetism or gravity. These two forces have the property that the potential[11] between two particles decays as $1/r$, and thus becomes weaker the longer the distance between the two particles is. Similarly, the force between two color-neutral particles, the residual effect of the strong interactions due to meson exchange, even decrease quicker because of the Yukawa potential (5.2) between them.

The situation for the interactions between two colored objects is drastically different. In fact, this was one of the reasons why it took very long to identify the underlying theory for the strong interaction, as its behavior is quite unique: The strong force becomes stronger with distance. An appropriate modeling of the inter-quark force is given by the potential

$$V = -\frac{4}{3}\frac{\alpha_s}{r} + \sigma r, \tag{5.12}$$

i. e. it has two components. There is a short-range Coulomb-like interaction, with the strong fine-structure constant $\alpha_s = g_s^2/(4\pi)$ defined by the strong coupling constant, similar to the electromagnetic force. But there is a second component, which rises linearly with distances, with a constant of proportionality σ. This latter contribution is of purely non-perturbative origin, which also made it very hard to discover its theoretical origin. That we do need such a potential for the strong interaction is an experimental fact. Only

[10] There are some subtleties associated with non-analyticities in this case, but in most cases they can be dealt with.

[11] Note for the following that the notion of potential has no real meaning anymore when particles can be created or destroyed, and it is necessary to pass to quantum field theory for a more precise description. But for the present purpose the analogy is sufficiently good as to give a qualitative correct picture, and is even quantitatively surprisingly good.

that α_s and σ are not unrelated can be understood so far, but why this is necessary is just an observation.

This explains why quarks appear to be lighter at high energies and thus shorter distances. At long distances, the interaction is very strong, which also manifests itself in chiral symmetry breaking and the condensation of quarks. There the quarks behave as described, and are quite heavy. However, this interaction becomes weaker at higher energies, i. e. quarks moving through the quark condensate at high enough energies will no longer be slowed down by it. The quarks will thus behave as if they only have their current mass. This explains the difference of both masses. The effect that the strong interaction becomes weaker at higher energies is called asymptotic freedom. Of course, the remaining force is still as strong as the electromagnetic one, due to the classical Coulomb potential in (5.12), but this is much weaker than at long distances, and therefore effectively free. Moreover, quantum corrections reduce this interaction further: The running strong coupling decreases at high energies, because the leading coefficient of the β-function β_0 in (3.11) is negative, with a scale of roughly 1 GeV. Hence, at a very high energy, the electromagnetic force becomes actually stronger than the strong one, above roughly $10^{15} - 10^{16}$ GeV.

The potential (5.12) also explains why no individual quarks are seen, at least to some extent. Since the potential energy between two quarks rises indefinitely when increasing the distance, one needs enormous amounts of energy when trying to separate them, much more than available for any earthbound experiment for any experimentally resolvable distance. This energy is stored in the gluonic field between the two quarks. Investigations have shown that this field is highly collimated between the two quarks, it is like a string connecting them both. Since the tension of this string is characterized to lowest order by the parameter σ in (5.12), this parameter is therefore also called string tension. This explains[12] why colored quarks are not observed individually, but are confined in hadrons, and the phenomenon is denoted as confinement.

However, in practical cases the energy limits of experiments is not a real concern. Long before the relevant energy limit is reached, there is so much energy stored in the string between the two quarks that it becomes energetically favored to convert it into hadrons. Hence, when trying to separate the two quarks inside, e. g., a meson by pumping energy into it, at some point it will just split into two mesons. This is called string breaking or screening, and is a 100% efficient mechanism, i. e. there is no remaining quark. The reason is that color, like electric charge, is conserved. Hence, overall the system must remain color-neutral. Thus, there would be at least two colored objects left over, which would again create immediately a string. This is also the mechanism why gluons are not

[12] Again, the true QFT version of this is much more subtle, and not yet fully understood.

observed. When trying to separate gluons from a hadron, again a string is created, which breaks and yields just additional hadrons.

5.9 Hadronic resonances

A consequence of the potential (5.12) is that for each hadronic state, there can be further excited states. These are the hadronic resonances, and exist for both baryons and mesons. E. g. the first excited state of the nucleon is the N(1440), with a mass of 1440 MeV, as the name indicates. In total, about 28, experimentally more or less well established, resonances have been observed for the nucleons, and likewise for the other hadrons.

An interesting observation, which can be understood on basis of the string picture discussed in the previous section, is the presence of so-called Regge trajectories. A Regge trajectory is that the mass squared of the resonances are roughly proportional to its total angular momentum, i. e. the combination of spin and orbital angular momentum. This has been experimentally observed before the inception of QCD, and has led to the first formulation of a string theory as a theory of the hadronic interactions. Due to consistency problems, and the lack of ability to describe other experimental observations, this approach of a quantized string has been dropped later in favor of QCD.

All hadronic resonances are unstable, and decay in various ways. Most quickly are those where the decay proceeds by splitting into hadrons, which usually involves the decay into a lower resonance and some pions. E. g. the N(1440) decays in about 50-75% of the cases into a proton or neutron and a pion, with a decay time of roughly $(200 \text{ MeV})^{-1}$. Some resonances can kinematically not decay into two hadrons, and they de-excite by emission of photons or other particles. This happens on a much longer time scale. Such states are often considered as excited states in contrast to the resonances. These are under the decay threshold into purely hadronic final states. The much longer decay time is of course due to the fact that the strong force is, indeed, the strongest one, and all processes subjected to it occur very quickly. The only exceptions are those processes close to, but above, the hadronic decay threshold. There kinematic reasons, the absence of a large phase space, statistically suppress decays[13].

The study of the decays of hadrons has been very useful in the understanding of the strong force. As will be seen later in section 5.12, today decays of hadronic resonances have also become an important window into the study of other processes.

[13]Quantum interference effects can occasionally modify those statements, but they are rather good guidelines for estimates of decays.

5.10 Glueballs, hybrids, tetraquarks, and pentaquarks

When studying the quark model and the possible color charges, it becomes quickly clear that three quarks or one quark and one anti-quark are not the only possibilities how to create color-neutral objects. Tetraquarks, made from two quarks and two anti-quarks, as well as pentaquarks, made from four quarks and one anti-quark, are equally possible. There is indeed no a-priori reason why such bound states should not exist. However, the are not yet any unequivocally observed tetraquarks or pentaquarks, though at least for tetraquarks there are by now substantial circumstantial evidences available. That an experimental observation is so complicated can be motivated theoretically by two arguments.

The first effect is mixing. E. g. for a tetraquark, it is almost always possible to construct a meson with the same quantum numbers, i. e. the same spin, parity, charge-parity and electric charge. There is also the possibility to construct equally well a dimeson molecule. One of the most infamous examples is the σ meson[14]. It is a light neutral meson, with quantum numbers compatible, e. g., with the states $\bar{u}u$, $\bar{d}d$, $\bar{u}u\bar{d}d$, $\bar{u}u\bar{u}u$, $\bar{d}d\bar{d}d$, $\pi^+\pi^-$, and $\pi^0\pi^0$. Since it is a quantum state, it follows the quantum rules, which in particular imply that all states with the same quantum numbers mix. It is therefore a superposition of all such states. The question which of these states contribute most is highly non-trivial. It can, in principle, be experimentally measured or theoretically calculated. There is no really reliable way of estimating it. The results found so far indicate that the combination of two pions is most dominant, it is therefore likely a molecule. For most other states the two-quark components appears to be the dominant one. Similarly, almost all baryons turn out to be completely dominated by the three-quark state. There are only few cases, where this may be different.

The second possibility is to investigate one of those possibilities where the quantum numbers of the tetraquarks cannot be created using a two-quark system. Such cases are rare, but they exist. In principle, it would therefore be sufficient to just observe such a state. Unfortunately, almost all of these states are highly unstable. They are therefore experimentally hard to observe, and it is thus challenging to establish their properties beyond doubt. Only very few candidates have been found so far, but some of them, to be discussed later in section 5.12, appear very promising.

This problem becomes more complicated due to the gluons. Though it is not possible to create a color-neutral state from a quark and a gluon, it is possible to combine a quark, an anti-quark and one or more gluons to obtain a colorless state. It is similarly possible to combine three quarks and a number of gluons to obtain a colorless state. Such states are

[14]The official name is $f_0(500)$, though the historical name of σ meson is still commonly used.

called hybrids. However, the gluons can at most add angular momentum, but no other charges to the state. Therefore, there is always a state with the same quantum numbers, but just made from quarks. Since adding a particle or orbital angular momentum to a state usually increases its mass, these states are unstable against the decay to a state with just the minimum number of quarks. Though these hybrids are therefore formally admixtures to any state, it is essentially always a small one, and therefore hybrids are very hard to identify both experimentally and theoretically.

The last class of states which can come into the mix are bosonic glueballs, which combine only gluons to a colorless object. The usual counting rules of the quark model do not apply to them, but as a rough estimate even the simplest state is made out of four gluons. Such states carry no electric charge, and most of them have the same J^{PC} quantum numbers as mesons, and therefore mix. However, there are some candidates, particularly the so-called $f(x)2$ mesons, with x around 1500 MeV, which appear to have a large admixture from glueballs. This is experimentally identified by the possible decays. Since gluons are, in contrast to quarks, electrically neutral, decays into electrically neutral decay products, except for photons, should be favored if there is a large glueball contribution in the state. This has been observed, especially when comparing the partial decay widths of decays to uncharged particles to the one to photons. However, even at best these state are only partially glueballs.

There are some glueball states which have quantum numbers which cannot be formed by only two or three quarks, at least not in the simple quark model. Unfortunately, all theoretical estimates place theses states at masses of 2.5-3 GeV, and therefore far above any hadronic decay threshold. They are therefore highly unstable, and decay quickly. Hence, there is not yet any experimental evidence for them, though new dedicated searches are ongoing or are prepared.

5.11 Flavor symmetry and strangeness

So far two quark flavors, up and down, have been introduced. As discussed in section 4.11, this implies the presence of an approximate SU(2) flavor symmetry, since the masses of both quarks are small. It was possible to describe all hadrons introduced so far using just these two quarks with the quark model.

However, already before QCD was formulated, hadrons were observed, which do not fit into this picture. The most well-known of them are the kaons $K^{\pm}$, K^0, and $\bar{K}^0$, four mesonic hadrons of masses 494 MeV for the charged ones and 498 MeV for the two uncharged ones. Most remarkably, these new mesons were more stable than those of similar masses made from the two quarks inside the quark model.

The resolution of this mystery is that there are more than the two quark flavors necessary to construct the proton and neutron. These additional quark flavors, which will be discussed now and in the following two sections, do not occur in naturally observed atomic nuclei[15] but can be produced in natural or artificial collisions.

The quark to obtain the kaons in the quark model has been called the strange quark s. Its current mass is 90-100 MeV, and therefore its full mass is about 400-450 MeV. It follows then that the kaons, with only 500 MeV mass, are also would-be Goldstone mesons, obtaining their mass from the, with the strange quark enlarged, chiral symmetry breaking. Since the strange quark's current mass is already not too small compared to usual hadronic scales, the kaons are much closer to their expected mass of about 700 MeV in the quark model obtained from a construction of a u or a d quark and one of the new s quarks.

Indeed, the s quark has an electric charge of $-1/3$, just like the d quark. The charged kaons are therefore the combinations $u\bar{s}$, $\bar{u}s$, and the neutral ones $d\bar{s}$ and $\bar{s}d$, explaining their small mass difference, and their multiplicity. The Goldstone theorem actually predicts that there should be 8 Goldstone bosons. These are six. The other two are the η and η' mesons, which are made from $\bar{s}s$ combinations, and some admixtures from neutral combinations of u and d quarks. They are therefore even heavier, the η having a mass of 550 MeV. Somewhat peculiar, the mass of the η' is much higher, about 960 MeV. The reason is that the η' also receives mass from another source, the so-called axial anomaly. The latter will be discussed below in section 6.11.

Besides the broken chiral symmetry, there is also the now enlarged flavor symmetry. Since it involves up, down, and strange quarks, it is an SU(3) group. Since the quarks have different masses, the group is not unbroken, but reduced to three counting symmetries, i. e. $\mathrm{U}(1)^3$. Hence, the individual quark flavors are conserved, but bound states with differing quark content have differing masses. This conservation of quark flavor by the strong interaction is also at the origin of the name strangeness. When the kaons were discovered, the quark model was yet to be established. The kaons, and also baryons, called hyperons, with a single or more strange quarks included, showed a different decay patterns than ordinary hadrons, due to the conservation of strangeness. Thus, they did not fit into the scheme, and were therefore considered strange.

The presence of the strange quark, which has an effective mass of about 400 MeV, and thus close to the masses of the up and down quarks of 300 MeV, adds many more possible combinations to the quark model, which all have very similar masses. Thus, there is a very

[15]With the possible exception of the inner core of neutron stars, though this is not yet settled.

large number of hadrons with masses between 500 MeV, and roughly 3000 MeV, where the states become too unstable to be still considered as real particles. In fact, the number of states N turns out to rise exponentially with mass, $N \sim \exp(M/T_H)$. This is a so-called Hagedorn spectrum, where T_H is called the Hagedorn temperature. The reason for the name is that naively a system with such a spectrum has the property that at a finite temperature, the Hagedorn temperature T_H, an infinite number of particles is created, and thus the system cannot be heated beyond this point. For the strong interactions, the Hagedorn temperature is about 160 MeV. Of course, it is in practice possible to go beyond this temperature. What happens is that at this point the quark substructure can no longer be ignored, and this effect limits the number of states growth. This has originally lead to the idea that at this temperature a phase transition has to occur, which signals the change from a hadronic system to one were the quark substructure becomes more important. However, the most reliable results to date rather indicate that it is a cross-over. This transition plays nonetheless a role in the development of the early universe, though a rather small one. This will be discussed in detail in section 5.18

The experimental and theoretical study of this transition is an important part of modern particle physics. Nowadays it is not restricted to the question of large temperatures, but also small temperature and large densities. This is the situation which is encountered during supernovas and in the interior of neutron stars and during neutron star mergers. At theses densities, which become larger than the density of nuclei, nucleons can no longer exist without starting to overlap with each other. Similarly to the high-temperature case, such an overlap of nucleons is expected to change the properties of matter dramatically, as the quark substructure can no longer be ignored. However, much less is known about this situation, as it is both experimentally and theoretically much more demanding to investigate. Some hope is provided by the possibility to use gravitational waves to probe the interior of neutron stars during neutron star mergers, but this will crucially depend on the developments in the next few years, whether this turns out to be correct. One of the most interesting questions is, whether strange quarks play a role for neutron stars.

Another aspect were strange quarks play an important role are the quests for tetraquarks and pentaquarks. Since strangeness is a conserved quantum number of the electromagnetic and strong interaction, it is possible to construct states which do not have the quantum numbers of an ordinary state, e. g. a meson with total strangeness of 2, which therefore must contain two strange quarks, and cannot be a simple quark-anti-quark state, or a baryon with strangeness -1, due to a single anti-strange quark, which therefore must be a tetra-quark. Searches for such signatures are intensively pursued. Note that, e. g., tetraquarks could also consist out of two strange and two antistrange quarks. In this case, the excess strangeness is only visible in the decays. Such states are therefore called cryptoexotic.

5.12 Charm, bottom and quarkonia

It appears at first rather surprising that there should be just one other quark, which has the same electric charge as the down quark. This appears unbalanced, and another quark with the electric charge of the up quark appears to be necessary. Indeed, this is correct, and there is also a heavier copy of the up quark, which is called charm[16] quark. However, while the strange quark has a rather similar mass, despite its larger current mass, as the light quarks, the charm quark has not. Its current mass is 1275 MeV, and thus similar to its effective mass of roughly 1600 MeV. As a consequence, hadrons involving a charm quark are much heavier than hadrons containing only the lighter quarks. Furthermore, though chiral symmetry is even more enlarged, its so badly explicitly broken that would-be Goldstone bosons involving charm quarks have essentially the same mass as without the breaking of chiral symmetry. The same is true for the flavor symmetry, and the only remnant is that charm is again a conserved quantum number in both the electromagnetic and strong interaction.

This conservation of charm has very interesting consequences. Of course there are hadrons were only one of the quarks is a charm quark, which are called open charm hadrons. The best known ones are the D mesons, with masses of about 1870 MeV mass and having the structure of a single charm quark and either an up or down quark. These are the lightest particles with a charm quark.

But there are also particles, especially mesons, which consists only of charm quarks. In the meson case, where the total charm is zero if they consist out of a charm and an anti-charm quark, these are said to have hidden charm and are called charmonia. The latter states are particularly interesting, because they show a very interesting mass spectrum. In fact, the lightest $\bar{c}c$ states have a mass which is below threshold for the decay into two hadrons with a charm quark and an anti-charm quark each, the $\bar{D}D$ threshold. They can therefore not decay directly. Of course, the charm and anti-charm quarks can annihilate. But because of how quark and gluon color charges are arranged, such a process is substantial suppressed in QCD compared to the decay with a production of an additional quark-anti-quark pair. The reason is the so-called Zweig rule. It can be understood in the

[16]The name originates from the fact that it solves several experimental mysteries observed in the weak interactions, to be discussed in chapter 6, and because at that time it appeared to complete the quark picture.

following way. When two quarks annihilate, this is a colorless state. Since gluons carry color, they cannot annihilate into a single gluon, though a single photon is possible. Two gluons is also not possible, since this state has not the correct quantum numbers in terms of the combination of spin, parity, and charge-parity. This is only possible starting with three gluons. But this requires three interactions. And though the interaction strength of the strong interaction is larger than for QED, three interactions is still rare, and therefore not a substantial effect.

Hence, decays occur very slowly. Thus, these hadrons are extremely stable compared to hadrons made from lighter hadrons, where the pions offer a simpler decay channel. These meta-stable charmonia states have therefore masses of about 3 GeV, but decay widths of around a few 100 keV.

Because of this fact, the charmonia states turn out to present a very good realization of the possible states permitted by the potential (5.12). Similarly to the hydrogen atom[17], this potential creates states distinguished by a main quantum number and orbital quantum numbers. The most well-known state is the J/Ψ, at about 3097 MeV with a decay width of 93 keV, which is a state with one unit of angular momentum. However, the ground state of the system is the η_c, with a mass of 2984 MeV and a decay width of 320 keV. That the ground state decays quicker is mainly due to kinematic effects from the angular momentum. Simply put, the ground state is in an s-wave, and thus the wave-functions of the two charm quarks have a large overlap. Thus an annihilation into photons is much more likely than in the case with angular momentum, where the overlap of the wave functions is much smaller. Right now about 8 states are known, which are below the $\bar{D}D$ threshold, the heaviest the so-called $\psi(2S)$, with a mass of 3690 MeV and a decay width of 303 keV.

These charmonia have been very instrumental in understanding the potential (5.12), and thus the strong interactions. The very well-defined spectrum, which provides the opportunity of a true spectroscopy, including many angular momentum states, permits a much cleaner study than in case of the light hadrons, where the ubiquitous decays into pseudo-Goldstone bosons make resonances decay very quickly.

However, not all of the states in this spectrum are easily explained within the framework of the quark model and the potential (5.12). These are the so-called X, Y, and Z mesons, with masses above the $\bar{D}D$ threshold, and some also with open charm. These states do not fit into the spectroscopic scheme, and especially some may have quantum numbers, which are not in agreement with a simple quark-anti-quark system. This is still under

[17] Spin-angular momentum couplings are much more relevant than for atoms, and have to be taken into account already at leading order.

experimental and theoretical investigation. However, it already shows that the simple quark model appears not to be able to explain all hadrons.

With the charm quark, it may appear that everything is complete and symmetric. However, nature did not decide to stop at the charm quark, but added another quark: The bottom (or in older texts beauty) quark. It is a heavier copy of the down quark, and has therefore an electric charge of $-1/3$. Its mass is about three times that of the charm quark, with a current mass of 4.2 GeV. It therefore introduces another quark flavor, but like for the charm quark, this does not play any role for dynamical chiral symmetric breaking, as the explicit breaking is far too large.

Other than the mass and the electric charge, the bottom quark behaves essentially as the charm quark. Especially, there is a rich spectroscopy with open and hidden beauty[18], the latter also called bottonium in analogy to charmonium. Similar to the case for the charm quark, the lightest mesons with open beauty are rather heavy, $B^{\pm}$ and B^0 at 5.3 GeV. As a consequence, the bottonium spectrum has a large number of quasi-stable states, the lightest being the η_b with a mass of 9.4 GeV and a decay width of roughly 10 MeV, the Υ playing the role of the J/ψ with a mass of 9.5 GeV and a width of 54 keV, and then even 15 states up to the heaviest $\chi_b(3P)$ with 10.5 GeV observed so far. There are also heavier states, including bottom versions of the X, Y, and Z mesons, which do not fit easily into a simple quark model explanation. Thus, an even richer spectroscopy is possible, though the production of bottonia in so-called beauty farms, requires more resources than for the charmonia.

Of course, for both bottom quarks and charm quarks there exist also baryons, with one or more of these quarks, also with both charm and bottom quarks. These are rather complicated to produce, but have been observed, though baryons with multiple charm or bottom quarks only very recently. These baryons are not as stable as the mesons, but are still sufficiently stable that their production and decay take place so far apart from each other, a few μm, that both processes can be experimentally resolved. They and the mesons play therefore an important role to identify the production of charm and especially bottom quarks in high-energy collisions (so called c- and b-tagging).

Together, charmonia and bottonia are usually referred to as quarkonia. Studying theses states are also interesting for other reasons than to understand QCD. Because the J/ψ and Υ are very long-lived, they are very well suited for precision measurements. Furthermore, as will become evident in section 6.5, it appears plausible that new phenomena will be influencing heavy quarks stronger than light quarks. Searching therefore for deviations of the decays of quarkonia, especially bottomonia, from the standard-model predictions has

[18] For the flavor quantum number the term beauty still survives.

become an important branch in experimental physics, which is also called flavor physics.

5.13 Top

With the introduction of the bottom quark the situation appears again as unbalanced as with the introduction of the strange quark. This is indeed true, and the picture is extended by the top (or in old texts truth) quark. This is the last quark, which so far has been found, though there is no a-priori reason[19] to expect that there may not be further quarks, and searching for them is a very active field in experimental physics.

The top quark is a heavy version of the up quark, and thus has an electric charge of 2/3. The fact which is really suprising about it is its mass of 173 GeV. It is therefore forty times heavier than the bottom quark, and this is the largest jump in masses in the quark mass hierarchy. It is an interesting side remark, and a triumph of theory, that this mass has been established within 10 GeV before the direct observation of the top quark, by measuring other processes sensitive to the top quark very precisely, and then using theory to make a prediction.

The enormous mass of the top quark makes it the heaviest particle detected so far. Due to its large mass, it decays much quicker than the lighter quarks, with a decay width of 2 GeV. This is a consequence of effects to be introduced in section 6. Hence, the formation even of short-lived hadrons with a top quark is almost impossible, and no hadron with a top quark has so far been directly observed. Whether a quasi-stable toponium is possible is not clear, but so far there is no experimental evidence for it. But, due to the mass, it is also not trivial to produce large numbers of top quarks, and thus the study of top quarks is rather complicated. Hence, the top quark is so far more of interest for its own properties, particularly its mass, rather than for its relation to QCD.

5.14 Partons and scattering experiments

Due to confinement, quarks (and gluons) cannot be isolated as individual particles. This even applies to the top quark though this is of little practical relevance as it decays so quickly. This has quite important consequences for the experimental detection of quarks and gluons, as it implies that only hadrons are ever observed.

---

[19]From the conceptual point of view. Within the standard model, such a fourth generation, given the lower mass bounds from expeirments, would yield so strong interactions with the Higgs of section 6.5 that it would be unlikely to have escaped detection.

This also implies that a hadron is, in fact, a complicated structure of the quarks providing its quantum numbers, the so-called valence quarks, and many virtual particles. The latter can be either also quarks, then called sea quarks to distinguish them from the valence quarks, and (also virtual) gluons. Collectively, the valence quarks and the virtual particles inside a hadron are called partons, a term originating from a time as it was not yet clear whether there are different particles inside a hadron, or just a single type. Furthermore, since quarks are indistinguishable, there are also processes where a sea quark and a valence quark are exchanged. Thus, valence quarks are not immutable.

The collective of all partons carries the total quantum numbers, like momentum, electric charge, spin etc., of a hadron. Especially, each parton carries a momentum fraction x (called Feynman x) of the total hadron momentum. If there would be only the valence quarks, each of them would have $x = 1/3$. Because of the virtual particles, the average x for each particle is smaller, since there are more particles, though quantum fluctuations permit also values of x larger than $1/3$, up to 1, though rarely. Since gluons are massless, they contribute the majority of particles at small x. As a consequence, many properties of hadrons can already reasonable well be approximated when the sea quarks are neglected, the so-called quenched approximation.

5.14.1 Fragmentation functions and jets

One of the most direct consequences of confinement is that the quarks and gluons are not observed. In fact, when they are produced in a high-energy collision they behave, due to asymptotic freedom, for a short time as quasi-free particles. But confinement kicks in after time or distance scales of about or less than a fermi. Since the initial state has been color neutral, the final state is also color neutral. Thus, any produced partons are connected, in the string picture by strings, with all the other colored particles. While the distances are small, the string is short and little energy is required. When, after the collision, the particles start to move apart, energy must be invested into the string, and it breaks eventually. This produces a spray of hadrons. It is said that the partons hadronize.

In the rest frame of the produced quark or gluon, this hadronization will be isotropic, if sufficiently many quarks and gluons are present that the color charge is roughly isotropically distributed. That is a rather good approximation, due to the light gluons, for most cases. But in high-energy collisions the quarks and gluons are usually produced at high energies. Hence, in the laboratory frame the produced hadrons will be highly collimated along the direction of the original quark and gluon. Such a collimated spray of hadrons originating from a single quark or gluon is called a jet.

This jet, in principle, carries all the information of the original quark and gluon, like electric charge, momentum, invariant mass, and flavor. In practice, some particles may still be not emitted very close to the original direction, i. e. not in the jet cone, and there may be other jets in an event. Thus, such information is not always easy to reconstruct. Still, by identifying such a jet, especially if it carries very much energy, it is possible to infer information about the original quark or gluon. E. g. top quarks are identified by the total invariant mass of the jet.

A further information can be obtained from which hadrons are present in the jet. Since quarks and gluons have different flavors, masses, spins, and electric charges, they produce different hadrons. The mathematical description of the probability for a hadron of type i to emerge from a quark or gluon of type j is called a fragmentation function f_{ij}. Such fragmentation functions are very hard to calculate, but can be measured experimentally, to some extent.

5.14.2 Parton distribution functions

Even more important in actual experiments is the inverse question: How probable is it to find inside the projectile hadrons (e. g. the protons at the LHC) a certain quark flavor or a gluon with a given momentum fraction x of the total hadron momentum? The answer to this question are given by the parton distribution functions $P_{hi}(x)$ for a hadron of type h and a parton of type i with momentum fraction x. The practical importance of these functions becomes immediately clear when considering some reactions which lead to a given final state f. The actual cross-section is then a convolution[20]

$$\sigma_{hh\to f} = \sum_{ij}\int dxdy\sigma_{ij\to f}P_{hi}(x)P_{hj}(y). \tag{5.13}$$

Hence, the actual cross-section gets contribution from all possible parton combinations inside the projectile hadrons.

The calculation of the parton distribution functions, or PDFs for short, is highly complicated. It is, especially for small values of x a non-perturbative question, and has not yet been answered satisfactorily. Hence, they are usually parametrized using experimental input, where theoretical considerations determine the parametrization. The experimental results are most easily obtained when instead of two hadrons electrons and hadrons are collided, as has been done at DESY in Hamburg with electrons on protons.

The importance of the PDFs cannot be underestimated, since they vary substantially with x. Especially, except for the valence quark flavors they drop very quickly with x.

[20] This is somewhat symbolic, and the indices i, j, and f include many further properties, like spins, polarizations etc.

As a consequence, it is rather unlikely to find a parton with large energy fraction inside a hadron. Since the cross-section for particle production are usually peaked around the mass of the particle to be produced, see equation (2.8), the effectiveness of producing new particles drops quicker with the energy than would be expected from just the energy dependency of the elementary cross section. Hence, the parton luminosities, i. e. the luminosity folded with the PDFs, is the actual figure of merit for a hadron collider. On the other hand, this means that increasing the energy for a hadron collider can increase the effectiveness faster than the naive energy scaling of the elementary cross-section.

PDFs have also a theoretical importance. Because of asymptotic freedom, it is often possible to obtain perturbatively a rather good approximation to the elementary cross-section $\sigma_{ij\to f}$. The involved PDFs are then known experimental input, and the interesting process can be obtained directly. That is the standard procedure used for most calculations in high-energy experiments. The assumption is that (5.13) is valid, i. e. that the process can be factorized into the PDFs and the elementary interaction. Because of asymptotic freedom, this is a reasonable assumption so far as the fractions x and y yield energy fractions for the corresponding partons much larger than 1 GeV. Since an integration is performed over all x and y, however, this will only work if the product of the PDFs and the cross-section becomes small at small x and y. For many cases of practical relevance, this is still a good approximation. Otherwise, so-called factorization violations are found, and this approach breaks down.

Formally, due to the presence of virtual particles, any elementary particle has to be described in terms of a PDF. E. g., an electron is surrounded by virtual photons. There is hence a PDF describing the probability to observe a photon with an energy fraction x of the total electron momentum inside an electron, $P_{e^-\gamma}(x)$. Of course, also quarks or gluons can be found in this virtual cloud. However, since the electron is not strongly interacting, the PDFs for non-partons can be often calculated to a reasonable accuracy perturbatively. Still, since especially at low energies also hadrons can be encountered inside an electron or a photon, this is not completely possible. This is not an esoteric effect. E. g. in the interaction of photons with nuclei the PDF $P_{\gamma\rho}(x)$ describing the mutability of photons into ρ mesons plays an important, and experimentally verified, role.

5.14.3 Monte-Carlo generators

As is visible, the experimental description of collisions is usually performed in a three step process, as far as factorization is viable. The first step is to obtain the initial particles for the interesting process. These are described by the various PDFs of the initial projectiles. The second step is to calculate the elementary process, usually performed perturbatively, to a certain order in perturbation theory. For the LHC at this time most such hard processes are calculated to orders between tree-level and NNLO. The third step is then to transform the final state of the elementary reaction into the particles finally observed in experiments. This involves usually first a decay cascade of the particles obtained in the elementary process into electrons, positrons, muons, photons, and other particles stable and weakly interacting on the scale of the experiments or into quarks and gluons. In the latter case also the creation of jets and/or fragmentation must be calculated, to obtain the final state.

While the initial state is rather simple, already the possibly involved intermediate particles before the hard collisions can easily number dozens. Moreover, even at moderate orders of perturbation theory, very many contributions appear. The final state is even more complex, and can involve many final particles, and several jets - up to eight jets and about a dozen quasi-stable particles are currently feasible for experiments.

To deal with this complexity requires computers. For many theories, the outer part of the event, the creation and fragmentation of the particles entering into the elementary hard process, is the same. Hence, this part can be automatized. Such programs, in which only the elementary cross-section is used as an input, are called Monte-Carlo generators. Current examples are publicly available codes like Sherpa, Pythia, Herwig, or Whizard, and can be downloaded e. g. at www.hepforge.org.

Even the elementary cross-sections can be calculated in many cases automatically, especially at tree-level. In these cases only the Lagrangian has to be provided, to permit an automated calculation from the initial to the final state.

5.15 Some algebra and group theory

Before continuing, it is at this time useful to interject a small introduction to a few more aspects of group theory. This will be useful to better understand the underlying structure of the difference of up and down quarks, as well as the properties of color in a formal theory later in section 5.16. Especially, the fact that three colors should create a neutral state, but not any other combination without anti-colors, already tells that it is impossible to do so just with real numbers. Vectors can indeed be arranged to have such a property. Since in two dimensions at most two vectors can be linearly independent, the space must be three-dimensional, to have three independent quark colors. But also the eight gluon colors, and the anti-colors must be fitted in. This requires more than just the standard mathematics used for electric charges.

It turns out that there is a unique mathematical structure which offers all these properties. This is the so-called Lie algebra su(3), and its associated Lie group SU(3). These are special cases of so-called Lie algebras and Lie groups. These are in turn special cases of symmetry groups and algebras, as introduced in section 4.1. In particle physics, some theory of these algebras and groups is inevitable, as it is not only necessary for the strong interactions, and QCD, but also for the weak interactions. In this brief section, the most basic concepts for Lie groups and algebras will be given. Much more group theory than the following is usually not necessary, though can be quite useful.

In fact, Lie groups are ubiquitous in particle physics. Examples have already been encountered in the context of flavor in section 4.11 and gauge symmetry 3.4, but there the details were not yet necessary.

To understand the concept, return to QED. QED implemented gauge-fields and matter which were related by the gauge-transformation to the u(1) algebra and the U(1) group. Thus, the fields could be regarded as tensor products of a function of space-time times an element of the group. To deal with QCD, it will turn out that it is only necessary to replace this u(1) by a different group, the aforementioned su(3), which is able to faithfully represent the concept of color. To proceed requires more insight into this group structure than for the flavor, and therefore some basic elements of the mathematical foundations of non-Abelian Lie algebras, which will be introduced before formulating QCD later.

Note that the fact that such mathematical structures describe particle physics is an assumption, which must be experimentally supported. This was indeed possible for the standard model of particle physics.

The basic element in this presentation will be to represent a quantity carrying a gluon-like color charge by the base vectors of a particular algebra. Such base vectors are called generators of the algebra. The type of algebra needed in particle physics are so-called Lie algebras, G. Hence, if there should be N independent gluon-like charges, there must be N independent base vectors τ^a with $a = 1...N$ and $N = \dim G$, and the Lie algebra must therefore be N-dimensional.

The defining property of such an algebra are the commutation relations

$$\left[\tau^a, \tau^b\right] = 2if^{ab}_{\ \ c}\tau^c \tag{5.14}$$

with the anti-symmetric structure constants f^{abc}. Note that sometimes the factor of 2 is included in the f^{abc}. This algebra implies already that everything is non-commutative, hence the name of non-Abelian Lie algebra. These structure constants fulfill the Jacobi identity

$$f^{abe}f^{cd}_{\ \ e} + f^{ace}f^{db}_{\ \ e} + f^{ade}f^{bc}_{\ \ e} = 0. \tag{5.15}$$

These base vectors have to be further hermitian, i. e., $\tau_a = \tau_a^\dagger$. Note that the position at top or bottom (covariant and contravariant) of the indices is of no relevance for the

Lie algebras to be encountered in the standard model, but can become important in more general settings: For the relevant Lie algebras, the metric is just δ^{ab}.

Such a Lie algebra can be represented, e. g., by a set of finite-dimensional matrices. An example is the su(2) algebra with its three generators, which can be chosen to be the Pauli matrices,

$$\begin{aligned}
\tau^1 &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \\
\tau^2 &= \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \\
\tau^3 &= \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.
\end{aligned}$$

Furthermore, to each algebra one or more groups can be associated by exponentiation, i. e.,

$$\lambda^a = e^{i\tau^a}, \tag{5.16}$$

provides base vectors for the associated group, which are by definition unitary and thus $\lambda_a^{-1} = \lambda_a^\dagger$. For su(2), these are again the Pauli matrices supplemented by the unit matrix, generating the group SU(2). However, the relation (5.16) is not necessarily a unique relation, and there can be more than one group representation. E. g., for su(2) there are two possible groups related to the algebra by the relation (5.16), the group SU(2) and the group SU(2)/Z_2 $\approx$SO(3), where matrices which differ only be a negative unit matrix are identified with each other.

Because of the exponential relation, a generic group element $\exp(i\alpha_a\tau^a)$ with real numbers α_a can be expanded for infinitesimal α_a as $1 + i\alpha_a\tau^a$. Thus the algebra describes infinitesimal transformations in the group. This will play an important role when introducing gauge transformations for non-Abelian gauge theories.

There is only a denumerable infinite number of groups which can be constructed in this way. One are the N-dimensional special unitary groups with algebra su(N), and the simplest group representation SU(N) of unitary, unimodular matrices. The second set are the symplectic algebras sp($2N$) which are transformations leaving a metric of alternating signature invariant, and thus are even-dimensional. Finally, there are the special orthogonal algebras so(N), known from conventional rotations. Besides these, there are five exceptional algebras g_2, f_4, and e_6, e_7, and e_8. The u(1) algebra of Maxwell theory fits also into this scheme, the u(1) group is the special case of all f^{abc} being zero, and the algebra being one-dimensional. This is equivalent to so(2).

The two algebras most relevant for the standard model are su(2) and su(3). The su(2) algebra has the total-antisymmetric Levi-Civita tensor as structure constant, $f^{abc} = \epsilon^{abc}$ with $\epsilon^{abc} = 1$. The algebra su(3) has as non-vanishing structure constants

$$
\begin{aligned}
f^{123} &= 1 \\
f^{458} = f^{678} &= \frac{\sqrt{3}}{2} \\
f^{147} = -f^{156} = f^{246} = f^{257} = f^{345} = -f^{367} &= \frac{1}{2},
\end{aligned}
$$

and the corresponding ones with permuted indices. There is some arbitrary normalization possible, and the values here are therefore conventional.

From these, also the generators for the eight-dimensional algebra su(3) can be constructed, the so-called Gell-Mann matrices,

$$
\begin{aligned}
\tau^1 &= \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \tau^2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \tau^3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\
\tau^4 &= \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad \tau^5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix} \quad \tau^6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \\
\tau^7 &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix} \quad \tau^8 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}.
\end{aligned} \tag{5.17}
$$

In general, there are $N^2 - 1$ base vectors for su(N), but the dependency for the other algebras is different. For the sake of simplicity, in the following only the expressions for su(N) will be given.

Generators, which are diagonal as matrices, and therefore commute with each other, are said to be in the Cartan sub-algebra or sub-group of the algebra or group, respectively. For su(2), this is only one generator, for su(3) there are two. This also defines the rank of the group.

These lowest-dimensional realization of the commutation relations is called the fundamental representations of the algebra or group. Since the commutation relations are invariant under unitary transformations, it is possible to select a particular convenient realization. Note, however, that there may be more than one unitarily inequivalent fundamental representation. E. g., for su(3) there are two, called fundamental and anti-fundamental. They are two linear independent realizations of the same group structure. For su(2), in contrast, there is only one such fundamental representation. There is always also the trivial representation, i. e., everything is mapped to the one.

It also possible to give representations of the algebras with higher-dimensional matrices, i. e. matrices which fulfill the commutation relations (5.14), but have a larger dimension. The next simple one is the so-called adjoint representations with the matrices

$$(A^a)_{ij} = -if^a_{ij},$$

which are three-dimensional for su(2) and eight-dimensional for su(3). There are cases in which the fundamental and the adjoint representation coincide. This can be continued to an infinite number of further representations, which will not be needed for the standard model, though they may appear in the context of beyond-the-standard-model physics.

The dimension of the representation can be loosely identified with the number of independent charges possible for a representation. Hence, the three quark colors require a three-dimensional representation, which is given by the fundamental one. The anti-charges require a further independent, three-dimensional one, which is given by the anti-fundamental one. The gluons have eight different colors. Thus, the representation needs to be eight-dimensional, which is the adjoint one. Hence, the assignment of representation is determined by the experimental fact how many different colors, or charges, there are. This also implies that if there is no suitable representation, the assignment of a group structure would be wrong.

This completes the required group theory necessary to describe the standard model, and most beyond-the-standard model scenarios.

5.16 Yang-Mills theory and QCD

It is now possible to return to QCD. It turns out that three quark colors (and three anti-quark colors) as well as eight gluon colors are precisely related to the group SU(3), i. e. the group of unitary three-by-three matrices with unit determinant. The three quarks can then be assigned to live in the corresponding three-dimensional complex vector space, the anti-colors in the corresponding hermitian conjugated space, and the eight gluon colors can be loosely associated with eight of the nine base matrices of SU(3), the ninth being the unit matrix. The eight non-trivial base matrices are the Gell-Mann matrices λ_a, given in (5.17).

More mathematically, following and generalizing section 3.4 for QED and using the group theory of section 5.15, the gluon fields live in the corresponding algebra, and are connected to the generators, and thus belong to the adjoint representation. The quarks and anti-quarks belong to the fundamental and anti-fundamental representation, respectively[21].

Is it thus sufficient to just generalize the QED Lagrangian (3.8)? The answer to this is both yes and no. The quarks are now vectors in color space, i. e. the spinors carry an additional color index like ψ^i. A gauge transformation acts now as a rotation in this color space, and is thus a matrix G as

$$\psi_i' = G_{ij}\psi_j = \left(\exp\left(i\frac{g}{2}\alpha_a\tau_a\right)\right)_{ij}\psi_j,$$

where the α_i are space-time-dependent functions, and g is playing the same role as the electric charge, and is thus a coupling constant. The corresponding covariant derivative to form a gauge-invariant Lagrangian has to take this matrix structure into account, yielding

$$D_\mu^{ij} = \delta^{ij}\partial_\mu - i\frac{g}{2}A_\mu^a(\tau^a)^{ij}, \tag{5.18}$$

where the A_μ^a are now the eight gluon fields. These have to transform under an infinitesimal gauge transformation like

$$A_\mu^{a'} = A_\mu^a + \frac{1}{g}D_\mu^{ab}\alpha^b, \tag{5.19}$$

$$D_\mu^{ab} = \delta^{ab}\partial_\mu - gf^{abc}A_\mu^c. \tag{5.20}$$

A finite transformation takes the form

$$i\tau^a A_\mu^{a'} = G\tau^a A_\mu^a G^{-1} + (\partial_\mu G)G^{-1},$$

were the combination $A_\mu^a\tau^a$ is often abbreviated as just A_μ. In the infinitesimal version (5.19) a second covariant derivative appears (5.20), where additional constants f^{abc} appear. These are just the structure constants of SU(3) from the defining algebra (5.14), and are therefore related to the Gell-Mann matrices as

$$\left[\tau^a, \tau^b\right] = 2if^{abc}\tau^c,$$

forming the algebra su(3).

Replacing the spinors by color vectors of spinors in the QED Lagrangian (3.9), and the flavors by the six quark flavors, is already sufficient to obtained the quark part of the QCD Lagrangian. However, it turns out that just taking the same field-strength tensor as in QED (3.7) with replacing the fields by their generalization with an index is not sufficient. In fact, in this case the Maxwell term (3.6) is not gauge-invariant.

[21]The actual group in the standard model is furthermore SU(3)/Z$_3$, that is all elements divided out which commute with every other elements, i. e. the center of the group called Z_3, except the unit matrix. This is of no relevance here.

This can be resolved by introducing a more complicated field strength tensor $F_{\mu\nu} = F^a_{\mu\nu}\tau^a$ with

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g f^{abc} A^b_\mu A^c_\nu. \tag{5.21}$$

It is a straightforward, though tedious, calculation to show that with this replacement the Maxwell-term (3.6) becomes gauge-invariant. However, in contrast to QED, this field-strength tensor in itself is still not gauge-invariant, but $\mathrm{tr} F_{\mu\nu}F^{\mu\nu}$ is. This can be seen by explicit calculation. For that it is useful to first establish the behavior of $F_{\mu\nu}$. Under an infinitesimal transformation, it follows that

$$\begin{aligned}
& \frac{1}{g}\partial_\mu D^{ab}_\nu \alpha^b - \partial_\nu D^{ab}_\mu \alpha^b + f^{abc}(A^b_\mu D^{cd}_\nu \alpha^d + A^c_\nu D^{bd}_\mu \alpha^d) \\
= & \frac{1}{g}(\partial_\mu\partial_\nu - \partial_\nu\partial_\mu)\alpha^a + f^{abc}(\partial_\mu A^c_\nu \alpha^b - \partial_\nu A^c_\mu \alpha^b) + f^{abc}(A^b_\mu \partial_\nu \alpha^c + A^c_\nu \partial_\mu \alpha^b) \\
& + g f^{abc}(f^{cde} A^b_\mu A^e_\nu \alpha^d + f^{bde} A^c_\nu A^e_\mu \alpha^d) \\
= & f^{abc} F^b_{\mu\nu} \alpha^c
\end{aligned}$$

as

$$\begin{aligned}
& (\partial_\mu\partial_\nu - \partial_\nu\partial_\mu)\alpha^a = 0 \\
& (\partial_\mu A^c_\nu - \partial_\nu A^c_\mu)\alpha^b + A^c_\nu \partial_\mu \alpha^b - A^c_\mu \partial_\nu \alpha^b + A^b_\mu \partial_\nu \alpha^c + A^c_\nu \partial_\mu \alpha^b = (\partial_\mu A^c_\nu - \partial_\nu A^c_\mu)\alpha^b \\
& f^{abc} f^{cde} A^b_\mu A^e_\nu \alpha^d + f^{abc} f^{bde} A^e_\mu A^c_\nu \alpha^d = (f^{adb} f^{ceb} + f^{aeb} f^{bdc}) A^d_\mu A^e_\nu \alpha^c \\
& = f^{abc} f^{deb} A^d_\mu A^e_\nu \alpha^c.
\end{aligned} \tag{5.22}$$

Here the anti-symmetry of the structure constants and the Jacobi identity (5.15) has been used. Note that it was essential that in (5.22) the coupling constant appearing in the gauge transformation and in the field strength tensor was identical.

Thus it remains the product of the two field strength tensors. To leading order, only one is transformed, yielding

$$f^{abc} F^b_{\mu\nu} F^{a\mu\nu} \alpha^c.$$

Since $F^a_{\mu\nu}$ is antisymmetric in the Lorentz indices, this vanishes, since the quadratic contraction is then symmetric in a and b.

QCD is therefore described by the Lagrangian

$$\mathcal{L} = -\frac{1}{4} F^a_{\mu\nu} F^{\mu\nu}_a + \sum_f \bar{\psi}^i_f \gamma_\mu (i D^\mu_{ij} - m_f) \psi^j_f. \tag{5.23}$$

This theory has a number of further remarkable features, which can be read off this Lagrangian.

The first is that the interaction between quarks and gluons is given by a vertex very similar to the QED one, it is just adorned by a Gell-Mann matrix. Since the interaction potential between quarks (5.12) is very different than the Coulomb potential, this implies that there must be substantial radiative corrections. In fact, since there is a qualitative change, these must be of non-perturbative origin.

Second, in QED it is in principle possible that each flavor has its own electric charge. This is not the case here. Due to the appearance of the second term in the field-strength tensor (5.21) the coupling constant in the gauge transformation in the gauge field (5.19) is fixed. But the covariant derivative (5.18) will then only yield the necessary cancellations between the transformations of the gluons and the quarks if the coupling constant appearing in it is the same as for the gluon fields. Thus, every flavor has to couple to the gluons with the same coupling constant g. This feature, which arises from the underlying group structure, is called coupling universality, and has been experimentally verified. It can also be seen directly from the Lagrangian.

Assume that there would be a difference, and the quarks would couple with q instead of g like the gluons, and that the gauge transformation for the quarks would be modified by some factor. Then

$$\begin{aligned}
& \left(\delta^{ij}\partial_\mu - i\frac{q}{2}A_\mu^{a'}\tau_{ij}^a\right)\left(\delta_{jk} + i\frac{p}{2}\tau_{jk}^b\alpha^b\right) \\
= \quad & \delta^{ik}\partial_\mu - \frac{iq}{2}A_\mu^a\tau_{jk}^a + \frac{ip}{2}\tau_{ik}^b(\partial_\mu\alpha^b + \alpha^b\partial_\mu) + \frac{pq}{4}A_\mu^a\tau_{ij}^a\tau^b jk\alpha^b - \frac{iq}{g^2}D_\mu^{ab}\tau_{jk}^a\alpha^b \\
\overset{p=\frac{q}{g}}{=} \quad & \delta^{ik}\partial_\mu - \frac{iq}{2}A_\mu^a\tau_{jk}^a + \frac{q^2}{4g}A_\mu^a\tau_{ij}^a\tau_{jk}^b\alpha^b - \frac{iq}{2}f^{abc}A_\mu^c\tau_{jk}^a\alpha^b + \frac{iq}{2g}\tau_{ik}^b\alpha^b\partial_\mu \\
\overset{q=g}{=} \quad & \delta^{ik}\partial_\mu - \frac{ig}{2}A_\mu^a\tau_{ik}^a + \frac{g}{2}A_\mu^a\tau_{ij}^b\tau_{jk}^a\alpha^b + i\tau_{ik}^b\alpha^b\partial_\mu \\
= \quad & \left(\delta_{ij} + \frac{i}{2}\tau_{ij}^b\alpha^b\right)\left(\delta_{jk}\partial_\mu - \frac{ig}{2}A_\mu^a\tau_{jk}^a\right)
\end{aligned}$$

Which confirms the statement.

Third, since the field strength tensor (5.21) is not gauge-invariant, so are not the chromoelectric and chromomagnetic fields, and then neither can be the chromo version of the Maxwell equation nor the color current. Hence, color and colored fields are not gauge-invariant, and therefore cannot be observable. Because of confinement, as described in section 5.8, this was not expected. But this makes confinement a necessity. Still, this leaves open the question why confinement operators in the way it does, with the potential (5.12), and not in any other way.

Fourth, the appearance of a quadratic term in the field-strength tensor (5.21) implies that there are cubic and quartic self-interactions of the gluons with each other. The coupling strengths, due to the underlying group structure, are uniquely fixed, again a consequence of coupling universality. But this also implies that gluons, in contrast to photons, are not ignorant of each other. A theory with only gluons in it is therefore not a trivial theory, but an interacting one. In fact, it turns out to be a strongly interacting one, exhibiting features like confinement or the formation of glueballs. This reduced version of QCD is called Yang-Mills theory. It is sometimes also denoted as quenched QCD, in the sense that any conversion of gluons in quarks (and thus sea quarks) is fully suppressed, i. e. quenched. It is possible to calculate in this reduced theory also approximately quantities like hadron masses. They are then found to be roughly of the same size as in (full) QCD, indicating that most of the physics in QCD resides in the gluon sector.

Fifth, a careful investigation of the vertices shows that the gluon-self-interaction and the quark-self-interaction are opposite. While quarks tend to reduce the interaction, i. e. they screen color charges, gluons tend to increase them, they anti-screen it. The experimental consequences of this fact actually lead originally to the discovery of the group structure of QCD, as this behavior is almost exclusive to theories with a non-Abelian Lie group as gauge group. Because anti-screening dominates this leads to the difference between QED and QCD, especially the feature of asymptotic freedom discussed in section 5.8. In the terms of section 3.8, the coupling does not become stronger but weaker with energies.

Especially, it is possible to calculate the β function (3.10) also of QCD. To leading perturbative order only the first Taylor coefficients is relevant, which takes for QCD the value

$$\beta_0 = \frac{11}{3}N_c - \frac{3}{2}N_f = 2,$$

where $N_c = 3$ is the number of colors and $N_f = 6$ is the number of flavors. The value is positive, in contrast to QED. Thus, because of the formula for the running coupling to leading order (3.11), the interaction strength in QCD diminishes with increasing energy, a manifestation of asymptotic freedom. The scale Λ turns out to be roughly 1 GeV, and the size of the coupling at about 2 GeV is 0.3, to be compared with the one of QED being about 0.0075. QCD is thus indeed much stronger interacting at every-day energies than QED.

However, at energies of the order of the proton mass, the perturbative coupling diverges, and a Landau pole arises. In this case, this clearly signals the breakdown of perturbation theory. In fact, even a value of 0.3 is already shaky for perturbative calculations, and strong deviations from perturbation theory have been encountered up to 25 GeV. Thus, only if all involved energy scales are in the range of (a few) 10 GeV or more QCD can be treated perturbatively.

However, it is possible to treat QCD at low energies also non-perturbatively. There are various methods available. The most successful one is arguably numerical simulations, which confirms that the QCD Lagrangian (5.23) describes objects like the proton and other hadrons. QCD is thus, to the best of our knowledge, the correct theory of hadrons, and hence of nuclear physics.

To obtain the full Lagrangian of QCD and QED, it is just necessary to add the QED Lagrangian to the QCD Lagrangian (5.23), and modify the covariant derivative of the quarks to include the electromagnetic interaction as

$$D_\mu^{ij} = \delta^{ij}\partial_\mu - igA_\mu^a(\lambda^a)^{ij} - ie_f A^\mu,$$

where now the electric charge depends on the flavor, and is either $2/3$ or $-1/3$ in units of the electron charge. This ratio of the electron to quark charge is actually very well established experimentally, though why it is the case is unclear. However, as will be discussed in section 6.11 it is necessary for the formulation of the standard model in its current form.

5.17 Cosmic radiation

Before adding further elements to the standard model it is worthwhile to discuss two examples where QCD (and partly also QED) plays an important role, which are usually not directly associated with particles physics.

One of them is cosmic radiation. There are many sources in the universe which create high-energetic particles. There is a wide range of such particles, e. g. leptons and photons, but also protons. Electrically neutral particles can be most easily tracked back to their origin, as their paths are not changed due to galactic or intergalactic magnetic fields, while charged ones are affected. These cosmic particles can have energies many orders of magnitude larger than created by any experiment. Energies of up to 10^9 TeV have been observed[22]. Possible sources of those most high-energetic particles are not yet conclus2ively identified, but may be active galactic nuclei.

When such a high-energetic particle hits earth's atmosphere, it reacts with the atoms and atomic nuclei. As a consequence, a jet, called in this context a cascade, of hadrons and electrons, muons, and photons is created, which can be observed with observatories on earth[23]. To be able to decide whether the original particle was charged or uncharged

[22]Such very high-energetic particles are only observed extremely rarely, at rates which can be as low as one per km^2 and year, and are thus not useful to test particle physics.

[23]Neutrinos and gravitational waves play by now also a role in astronomoy, but they will be explored in different chapters.

requires to understand the evolution of this jet. This is a complicated problem, as it requires to understand how the various particles interact and are produced. But this is decisively important, so that it becomes possible to understand whether the direction of the cascade could have been influenced by magnetic fields, and thus whether its origin is meaningful or not.

Similar to the Monte Carlo generators of section 5.14.3 for particle collisions, this can be usually only done by simulations. This is an important task for astrophysical observation, and our knowledge of the universe.

5.18 The QCD phase diagram

As has already been noted in section 5.11, aspects of QCD play an important role for the physics of neutron stars. In fact, a neutron star can be considered as a gigantic, stable nucleus, though this is an oversimplification. Since the density inside a neutron star increases from the outside to the inside, the physics will change. While the outer part is indeed essentially a nuclear system, the situation in the interior is not clear. Because QCD is so strongly interacting, it is very hard to treat in such an environment, and even numerical simulations fail[24]. It is therefore still an open field what actually occurs near the core of a neutron star, and whether the state of matter there is still just nuclei, or whether other hadrons, including strange ones, play a significant role.

The situation in neutron stars is only a part of a wider topic, the QCD phase diagram, i. e. what is the state of matter at densities of similar or larger size than in nuclei and temperatures of order of the hadron masses.

Experimentally, these are very hard to address questions. For neutron stars, possibly gravitational wave astronomy, as remarked in section 5.11, together with the X-ray spectrum of neutron stars as a function of time, as well as their sizes and masses, will be the only available experimental input for a long time to come.

At much smaller densities, the situation becomes essentially the one of nuclear physics. here, experiments with nuclei, especially collisions, can be used. In this way, it was possible to find that there is a phase transition between a gas of nucleons and (meta-)stable nuclei at roughly the density of nuclei, which permits to form nuclei. This phase separation

[24] The reason is actually not the strongness of QCD. It is rather a problem which has to do with the distinguishability of particles and anti-particles, and how this can be treated if there are more particles than anti-particles. For some theories, like QCD, but there are also such systems in nuclear physics and solid-state physics, this entails algorithmic problems, the so-called sign-problem, which prevents so far the development of efficient simulation algorithms, as the computation time scales exponentially with the system size. It is a technical problem, and not a physics problem.

persist for a few MeV in temperature, up to about 15 MeV, where it ends in a critical end-point. Thus, both phases are not qualitatively distinguished, just like in the case of gaseous and liquid water.

From the point of view of particle physics, the involved energy scales of nuclei are very small, and the distinction between the liquid and gaseous phases is essentially irrelevant. Thus both phases are not regarded as different. This common phase is denoted as the hadronic or vacuum phase, the latter as for all practical purposes the phase consists of far separated hadrons in a vacuum.

There are now several interesting directions to move on. Usually, they are signified by temperature and the baryon-chemical potential, i. e. the chemical potential distinguishing baryons and anti-baryons. Thus, zero (baryo-)chemical potential[25] denotes a situation in which there is the same number of baryons and anti-baryons, which includes the possibility of none of either type.

When neutron stars are interesting, the axis with zero temperature and finite baryo-chemical potential is most interesting. Neutron stars do have some temperature, but it is of the order of one MeV, and even during their formation in a supernova explosion or during a merger it does not exceed 10-20 MeV. On hadronic scales, this is essentially negligible. Thus, to good approximation it is a movement only along the baryo-chemical potential direction. Starting from the vacuum, i. e. zero baryo-chemical potential, it can be shown exactly that nothing will happen before reaching a chemical potential of roughly a third of the nucleon mass. This is due to some analyticity properties of the free energy as a function of the masses of the lightest baryon, and the fact that it is made out of three quarks. This feature is called the silver-blaze feature. However, any small temperature will change this, and it is then only approximately true for chemical potentials much larger than the temperature, but smaller than this silver-blaze point.

After this point soon the nuclear liquid-gas transition is encountered. After this experimentally established point, as noted above, no fully reliable results are available. There are some reasons to believe that at least one further phase transition will be encountered, though even this is not sure. It is furthermore unclear whether this will happen at densities sill relevant for a neutron star, or significantly above it. However, model calculations, i. e. calculations using simplified versions of QCD, as well as comparisons to other theories

[25]As in most cases only the baryo-chemical potential is present, the word baryo is dropped. The only other chemical potentials of relevance in most cases are an isospin-chemical potential, which gives the difference between up and down quarks, and is therefore relevant in neutron stars, though it is even then only small because of the close similarity of up and down quark, strange-chemical potential, which determines how many more strange than anti-strange quarks appear, and electro-chemical potential, which indicates the presence of electrons.

which are similar to QCD, indicate that there could even exist many different phases, some amorphous, and some crystalline, in which besides the nucleons also other hadrons, like pions, kaons, and hyperons, may play a role.

The situation is much better regarding zero chemical potential. This situation is relevant in the early universe, where though all matter in the universe is already present, and there is thus a sizable amount of baryons, the temperature is high enough to thermally produce baryon-anti-baryon pairs, reducing the chemical potential very close to zero.

This situation is good accessible experimentally by high-energy heavy-ion collisions, e. g. at the LHC with up to 2.4 TeV kinetic energy per nucleon for lead nuclei. In such an experimental setup, temperatures as high as 600-700 MeV with almost zero chemical potential, despite the 416 nucleons in the original nuclei, can be achieved. Furthermore, this situation poses no serious problems to numerical simulations. Hence, the knowledge of this axis is rather good.

It turns out that the physics depends significantly on the masses of the up and down quarks. Though this is also suspected for the remainder of the phase diagram, it is evident in this case. If both quarks are very heavy, there is a first-order phase transition at a temperature of about 250-300 MeV. As the quark masses become lighter, this temperature decreases, and the transition becomes a rapid cross-over at a temperature[26] of about 150-160 MeV, the aforementioned Hagedorn temperature. This is the situation for up and down quark masses observed in nature, the physical masses. The mass of the strange quark influences the precise values of the temperatures, but does not provide any qualitative influence, and the heavier quarks have even less relevance. If the quark masses are decreased further, the temperature still drops a little bit. More importantly, at some point the cross-over turns again into a phase transition, this time of second order and remains until zero quark masses.

What happens can be understood already in a simple picture. Temperature is classically nothing more than the kinetic energy of particles. In a quantum theory, temperature is just energy, which can also be converted to new particles. This will be exponentially suppressed with the mass of the created particles. Hence, the lightest particles will be most copiously produced. In QCD, these are the pions. These particles will have large kinetic energies, and will rapidly and repeatedly collide. Because of the asymptotic freedom of QCD, these scatterings will mainly be dominated by hard partonic scatterings, and are thus almost perturbative. Thus, QCD becomes essentially as it behaves at high-energies. Especially, this implies that the effects of chiral symmetry breaking become reduced, and

[26]There is no unique definition of a cross-over temperature. This is just the temperature were most changes occur.

the quarks lose their effective mass, though not their current mass, at the phase transition or cross-over. In fact, in the limit of zero quark mass, the second order transition becomes a symmetry transition where chiral symmetry becomes restored. At the same time, since most collisions are hard and partonic, excited states become very unstable and in most cases it does not matter anymore that quarks are confined into hadrons. They act effectively as if they would no longer be confined. Thus, one speaks also of a deconfined phase, and calls the transition a deconfinement transition. However, since it is a cross-over, it is clear that qualitatively nothing has changed, but quantitatively it is a completely different situation.

As a consequence of this dominance of the partonic degrees of freedom and asymptotic freedom actually the high-temperature thermodynamic behavior of the theory is essentially that of a free gas of quarks and gluons, a so-called Stefan-Boltzmann gas. The reason is mainly that the hard processes contribute to the free energy like the temperature to the fourth power, while all other effects contribute like the cube of the temperature, or even less. Hence, they become for thermodynamic bulk properties, i. e. extensive properties, irrelevant. Still, there are certain observables which are sensitive to non-trivial effects. Furthermore, the transition is very slow, and even at a few times the transition temperature even the bulk quantities are not yet fully dominated by the partonic processes. The system then behaves not like a gas, but rather like an almost ideal fluid.

This is also the situation encountered in the early universe. While it cools done, it will go through this cross-over. Before that, it is essentially dominated by the quarks and gluons, and only afterwards it starts to be dominated by the hadrons. However, because the transition is a cross-over, it seems that the transition had little quantitative influence on the evolution of the universe. Still, the point where it became possible to form stable nucleons is an important point, as this fixed the relative abundances of elements in the early universe. This process is called nucleosynthesis. The relative amount of nuclei created at this, essentially only hydrogen, helium, lithium, and their isotopes, have been both observed and calculated. Both theory and experiment agree for most isotopes rather well.

The situation in the remainder of the phase diagram is not yet clear. It is possible to map out parts of it with heavy-ion collisions at lower energies. Because then less energy is available, less particles are produced, and therefore the baryon chemical potential is larger. Still, the accessible region is that of rather high temperatures, above those characteristic for neutron stars, and likely below the relevant chemical potentials. Also, numerical simulations start again to fail the larger the chemical potential becomes. Hence, the situation becomes less and less clear. What seems to be certain at the current time is that for quite some distance into the chemical potential direction little changes, and the cross-over remains at a temperature only slowly decreasing with increasing chemical potential. There are some speculations about a critical end-point, from which a phase boundary starts, which eventually meets with the chemical potential axis, but this is not yet settled. Other than that, the field is still wide open.

Chapter 6

Weak interactions

The last ingredient of the standard model are the weak interactions, which will again be a gauge interaction, though of quite different character than both the electromagnetic and the strong interaction. The name weak is also somewhat misleading, as the interaction is actually stronger than the electromagnetic one, but it is short-range, as is the strong interaction, though for entirely different reasons.

It is impossible to separate from any description of the weak interactions the recently established Higgs sector, as without it the weak interactions would be very different in character. In fact, without the Higgs sector, it would be very similar to the strong interaction. The Higgs sector introduces the only non-gauge interactions into the standard-model. Depending on personal taste, the interactions of the Higgs are not counted at all as true interactions, as many believe these are just low-energy effective realizations of a yet unknown gauge interaction, a single interaction, or, by counting the number of independent coupling constants, 13 interactions. This will become clearer in the following.

One of of the most interesting facts about the weak interactions is that they live on a quite different scale than electromagnetism and the strong interactions. While electromagnetism acts mainly on the scales of atomic physics, and the strong interactions at the scale of nuclear and hadronic physics, the typical scale of the weak interactions will be found to be of the order of 100 GeV. Since this is the highest known scale in particle physics yet, apart from the gravitational scale, it is widely believed that the study of the weak interaction will be the gateway to understand whether there is any other physics between this scale and gravity. This is the main reason it plays such a central role in modern particle physics.

One of the reasons why the weak interactions, or more commonly denoted as the electroweak sector for reasons to become clear later, is rather unwieldy is that it is an aggregation of several, intertwined phenomena. Though each of them can be studied on its own, essentially none of them unfolds its full relevance without the others. This makes the electroweak sector often at first somewhat confusing. The different elements appearing will be a new gauge interaction, a mixing between this new interaction and QED, parity violation, the Higgs effect, and flavor changes.

6.1 β decay and parity violation

The weak interaction surfaced for the first time in certain decays, particularly the β-decay of free neutrons and of bound nuclei, which could not be explained with either the electromagnetic force nor the then newly discovered strong force. That this is impossible is directly seen from the quark content of the proton and the neutron: A transmutation of an up to a down quark is not possible in either interaction, because both are flavor-conserving. Furthermore, no hadrons are produced in the process, so this cannot be an exchange reaction. hence, another, third force has to be involved. Due to the small rate and also the very rarely seen consequences of this force in ordinary physics, this force was called weak force.

Though no additional hadrons are produced, other particles are involved. Especially the decay conserves electric charge by adding an electron to the final state, i. e. $n \to p+e^-$. However, for a free neutron this violates spin conservation. Furthermore, it is found that both the proton and the electron exhibit a continuous energy spectrum, and therefore energy conservation would be violated. The resolution of this is that the decay is indeed into three particles, and involves a new particle, the so-called electron-neutrino ν_e, actually an anti-electron-neutrino $\bar{\nu}_e$, $n \to p+e^- + \bar{\nu}_e$. Since the neutrino turns out to be a fermion of spin 1/2, this fixes both spin and energy conservation. The neutrino itself is a quite mysterious particles, and will be discussed further in section 6.2.

The first attempts of a field-theoretical formulation were based on a four-fermion interaction of type

$$G_F(\bar{p}\gamma_\mu n)(\bar{e}\gamma^\mu \nu_e), \tag{6.1}$$

where p, n, e, and ν_e represent the fields of the involved proton, neutron, electron, and the (anti)neutrino and are four-component spinors. The characteristic scale for the weak process was set by the Fermi constant, which is of order 1.14×10^{-5} GeV^{-2}. This is not a renormalizable interaction, and as such should be only the low-energy limit of an underlying renormalizable theory.

Besides the appearance of a new particle, the β-decay shows another quite surprising property: It violates parity. Thus, a β-decay is not operating in the same way as it would in the mirror. Experimentally, this can be observed by embedding the process into a magnetic field. A parity-conserving interaction, like electromagnetism, should yield the same result, irrespective of whether the spins are aligned or anti-aligned with the magnetic field[1], due to parity invariance. However, experimentally a difference is observed, indicating parity violation. More precisely, it was observed that a polarized neutron which decays will emit the electron preferentially in one direction. Therefore, the interaction must couple spin s and momenta, and would therefore have a contribution proportional to sp. However, the momenta of the decay products also depend on the invariant mass, p^2, and thus on a scalar contribution. Since therefore both scalars (scalar products of two vectors or two axial vectors) and pseudoscalars (products of a vector and an axial vector) appear imply that the interaction is not having a definite transformation behavior under parity, and is thus parity violating. In fact, it turned out that it is maximally parity violating.

To give this a more formal version, it is necessary to consider how couplings to fermions transform under parity. The parity transformation of a spinor is obtained by multiplying it with γ_0, the time-Dirac matrix. So, for spinors ψ the parity transformation is given by

$$P\psi = \gamma_0\psi.$$

Since furthermore $\gamma_0\gamma_\mu\gamma_0 = -\gamma_\mu$, the four-fermion coupling (6.1) will indeed transform as a scalar

$$P((\bar{\psi}\gamma_\mu\psi)(\bar{\psi}\gamma^\mu\psi)) = (\bar{\psi}\gamma_0\gamma_\mu\gamma_0\psi)(\bar{\psi}\gamma_0\gamma^\mu\gamma_0\psi) = (\bar{\psi}\gamma_\mu\psi)(\bar{\psi}\gamma^\mu\psi).$$

To obtain a pseudoscalar coupling, one of the vectors would have to be replaced by an axial vector. This can be obtained if there would be a matrix such that $\gamma_0\gamma_5\gamma_0 = \gamma_5$. In fact, such a matrix is given by

$$\gamma_5 = i\gamma^0\gamma^1\gamma^2\gamma^3.$$

This matrix anticommutes with all Dirac matrices,

$$\{\gamma_5, \gamma_\mu\} = 0.$$

As a consequence, the current

$$\bar{\psi}\gamma_5\gamma_\mu\psi,$$

is an axial vector, and can be used to obtain a pseudoscalar coupling.

However, this does not yet determine to which extent the weak interactions should be parity violating, and this can also not be predicted on basis of the standard model to be developed. Experimentally, however, it is found that parity is maximally violated. For massless particles (e. g. for neutrinos to a very good accuracy) this would imply that only

[1]In a practical case, the spin of the originating nucleus plays also a role, and a suitable choice is mandatory to see the effect simply. It was first observed in the β-decay of ^{60}Co.

one of the helicity states would be affected by the weak interactions. The helicity state for a spinor is projected out as

$$\frac{1\pm\gamma_5}{2}\psi.$$

The sign is determined by whether left-handed or right-handed states should be selected. Experiment finds that only left-handed states are involved, and thus a minus sign is appropriate. Furthermore, the weak interactions are found to violate also the charge conjugation symmetry maximally. Hence, the sign is not reversed for the anti-particle state. Therefore, the correct four-fermion interaction version of the weak interactions would be (appropriately normalized)

$$\frac{G_F}{\sqrt{2}}\left(\bar{\psi}\frac{1-\gamma_5}{2}\gamma_\mu\frac{1-\gamma_5}{2}\psi\right)\left(\bar{\psi}\frac{1-\gamma_5}{2}\gamma^\mu\frac{1-\gamma_5}{2}\psi\right),$$

which therefore exhibits maximal violation of C and P individually, but conserves CP.

Hence, the new interaction is very different from either the strong or electromagnetic force: There are particles only responding to it, the neutrinos, and it violates both parity and flavor conservation. Furthermore, it is found to be very short range, much more than the strong interaction. On the other hand, electrons and neutrinos are both sensitive to it, but are not confined. Thus, the way how this range limitation is operating has to be different from the confinement mechanism of QCD.

6.2 Neutrinos

Before discussing the weak interaction itself. It is worthwhile to introduce the neutrinos first. Some further properties of them will be added later, in section 6.9, when the weak force has been discussed in some more detail.

So far, only one neutrino has been introduced, the electron-neutrino, a fermion of spin 1/2. There are actually two more flavors of neutrinos, the muon-neutrino ν_μ and the τ-neutrino ν_τ, which are again fermions of spin 1/2. As their name suggest, they are partnered with one lepton each, forming the last member of the particle generations, which are conventionally assigned to be the sets $\{u,d,e,\nu_e\}$, $\{c,s,\mu,\nu_\mu\}$, and $\{t,b,\tau,\nu_\tau\}$. These combinations are somewhat arbitrary, and mainly motivated by the masses of the particles. Another motivation is that flavor is, though violated, still approximately conserved for leptons and neutrinos. It then turns out that, if electrons carry an electron flavor number of one and positrons of -1, electro-neutrinos likewise carry an electron flavor number of 1 and anti-electron neutrinos of -1. Hence, e. g. the β-decay conserves this flavor number. Likewise, there are approximately conserved flavors for muon-like and tauon-like particles.

As a consequence, neutrinos are usually also included in the term leptons, though not always.

When it comes to masses, the neutrinos actually are very surprising. To date, it is only known for sure that their mass is below[2] 2 eV. A direct measurement of the mass of the neutrinos is complicated by the fact that they are only interacting through the weak force, which reduces cross-sections by orders of magnitude compared to the leptons and the quarks. However, it was possible to measure the mass differences of the neutrinos in a way to be explained in section 6.9. But so far, it was only possible to measure differences without knowing which of the possible differences are measured. The two values are about 0.009 eV and 0.05 eV. This implies that at least one of the neutrinos has a mass of 0.05 eV, though it is not clear, which of them. It is furthermore not yet clear, whether the other two or just one is much lighter. If the electron neutrino would be the heaviest, there are very good chances to measure its mass directly within the next decade. Since the other neutrino species are much harder to produce in larger numbers, this will take much longer if it is is either of the other ones. It is also, in principle, still possible that one of the neutrinos could be massless, though there is yet no strong evidence for this possibility.

6.3 Flavor-changing currents

As already observed, the weak interaction changes the flavor of down quarks to up quarks. Likewise, processes have been observed that change strange to charm and bottom to top, and vice versa, kinematics always permitting. There are also processes observed which change, say, charm to up quarks. However, these processes are strongly suppressed, and will be neglected for now. They will be added below in section 6.8.

After neglecting these processes, the flavor changes only occur within one generation. Since in this process also one unit of electric charge is transferred to leptons, these are called flavor-changing charged current, or FCCC for short. In addition, there are also flavor-changing neutral currents, FCNC, but they belong to the class of processes postponed.

This observation strongly suggest that the quark flavors in any generation can be regarded just as two states, a doublet, under the weak interaction, similar to the isospin. Hence, a new symmetry is introduced, the weak isospin symmetry, under which doublets are formed from the two quarks of each generation, and also the lepton and the neutrino of each generation. This doublet structure has been already observed with the isospin relation (5.4), and it will be seen later how the electromagnetic charge enters this. Note,

[2]This is a conservative upper bound. There are good indications that it is at least an order of magnitude smaller.

however, that in FCCC also the masses play a role, and for reasons to be discussed later massless particles would not participate. Therefore flavor changes for the leptons occur much more rarely because of the very small neutrino masses, though they have been observed nonetheless.

6.4 W and Z bosons

The doublet structure can be regarded as a charge, the weak isospin charge. The weak force couples to this charge. The interaction which mediates the neutron decay appears to be described by the interaction of four fermions, an up quark, a down quark, an electron, and an anti-electron neutrino, as given in (6.1). The strength of this interaction is characterized by a new coupling constant, the Fermi constant G_F, which has units of inverse energy squared, and a value of 1.16×10^{-5} GeV^2, corresponding to an energy scale of about 293 GeV, though usually this energy scale is reduced by a factor of $2^{\frac{1}{4}}$ to 246 GeV for reasons of conventions. At any rate, the energy scale is much larger than any hadronic scale. The only other object encountered so far with a similar scale is the top quark, though this appears at the current time unrelated.

Since weak charges are not confined, another mechanism is required to give the weak interaction a finite range. According to section 5.1, a possibility is that the charge carriers are massive. Another problem with this theory was that such a theory is not renormalizable, in the sense of section 3.7. Hence, it would not be possible to obtain meaningful predictions from it at energies of similar size or larger than 246 GeV, which would already cover any top quark physics. It was therefore early on argued that this cannot be the underlying theory, and at a scale of around 100 GeV there must be a different theory present.

This indeed turns out to be true. In fact, it was suggested that there should exist, similarly to QED, an exchange boson. However, to produce such a scale and range, they would have to be massive with a mass M_W. Assuming a coupling constant g' for the process, the scale would be set by expressions of the type

$$G_F \approx \frac{g'^2}{M_W^2}, \tag{6.2}$$

indicating a mass scale of about 100 GeV for the process, if g' is not too different from one. This already sets the scale of the weak interactions. Furthermore, the appearance of a mass-scale indicates that the weak interactions will have only a limited range, of the order of $1/M_W$, and would be screened beyond this scale. Its range is therefore much shorter than that of any other force.

Over the time other weak effects were found, all characterized by the scale G_F. In particular, after some time the postulated interaction bosons have been observed directly. There are two charged ones, the $W^\pm$, and a neutral one Z^0, with masses about 80 and 91 GeV, respectively, and where the superscripts indicate their electric charge. In fact, the weak interaction turns out to have a very similar structure as QED and QCD in the sense that it is a gauge theory, where additional gauge bosons appear. The doublet structure already suggests as the underlying gauge group one with two different charges and three gauge bosons. It finally turns out that the correct gauge group is SU(2). That the $W^\pm$ carry electric charge indicate a relation to QED, which will be discussed in section 6.7.

However, masses are not permitted for gauge bosons, or else gauge invariance would be broken. This breaking has severe effects, making the theory non-renormalizable. Therefore, despite the experimental verification of the existence of the weak gauge bosons and their masses, this would not improve the theoretical description. This was only achieved after gauge theories have been combined with the Goldstone theorem described in section 5.7.

To understand that this is the right way, consider a photon A_μ coupled to some external current j in the Lagrangian

$$\mathcal{L} = -\frac{1}{4}(\partial_\mu A_\nu - \partial_\nu A_\mu)(\partial^\mu A^\nu - \partial^\nu A^\mu) - j^\mu A_\mu.$$

It will be this current, and its generalization, which will be instrumental in obtaining a massive gauge boson and in addition a gauge symmetry, and thus a superficially massive vector boson. However, since these two concepts are usually contradictory, this will lead to a hiding of the gauge symmetry, similarly as the global symmetry has been hidden in section 5.7.

The corresponding equation of motion for the photon is

$$\partial^2 A^\nu - \partial^\nu \partial_\mu A^\mu = j^\nu. \tag{6.3}$$

Furthermore, the field should, somehow, satisfy the condition

$$(\partial^2 + M^2)A^\nu = 0. \tag{6.4}$$

To proceed further, it is useful to choose a particular gauge, in this case the Landau-Lorentz gauge $\partial_\mu A^\mu = 0$. Of course, fixing a gauge is a perfectly acceptable way to perform a calculation, provided the calculation is not violating gauge invariance at any point. Then the results for gauge-invariant quantities, like the number of physical polarizations, will be valid. In the present case, doing a calculation without fixing the gauge is at best tedious. The results can be translated to the general case of an arbitrary gauge, but it should be kept in mind that any relation on gauge-variant quantities, like the gauge field A_μ, will only be valid in this particular gauge. However, statements about gauge-invariant quantities, like, e. g., the field-strength tensor $F_{\mu\nu}$ in this Abelian example, will hold in any gauge.

After imposing this gauge, the equation of motion (6.3) for the spatial components becomes

$$\partial^2 \vec{A} = -\vec{j}. \tag{6.5}$$

To simplify matters further, restrict to a time-independent situation. Then all time-derivatives vanish, and the equation (6.4) takes the form

$$\partial^2 \vec{A} = M^2 \vec{A}. \tag{6.6}$$

From the equations (6.5) and (6.6) it is clear that the current must satisfy the condition

$$\vec{j} = -M^2 \vec{A}$$

in order that the photon (in this gauge) becomes apparently massive and at the same time the gauge invariance is not broken. It is left to find some way of producing the appropriate current. Physically, the origin of such a current being proportional to $\vec{A}$ is due to the response of a medium to the acting electromagnetic fields. E. g., this is realized by the Meissner effect in superconductivity. Therefore, giving a mass to the photon requires a medium which provides a response such that the photon becomes damped and can therefore propagate only over a finite distance. In the electroweak case, the role of this medium will be taken by a condensate of Higgs particles.

This is a screening process as can be seen from the following simple example. Consider the Maxwell-equation

$$\vec{\partial} \times B = \vec{j} = -M^2 \vec{A}.$$

Taking the curl on both sides and using $\vec{\partial} \times \vec{A} = \vec{B}$ yields

$$\partial^2 \vec{B} = M^2 \vec{B}.$$

In a one-dimensional setting this becomes

$$\frac{d^2}{dx^2} B = M^2 B,$$

which is immediately solved by

$$B = \exp(-Mx).$$

Thus the magnetic field is damped on a characteristic length $1/M$, the screening length. The inverse of the screening length, being a mass in natural units, is therefore the characteristic energy scale, or mass, of the damping process. However, by this only an effective mass has been obtained. The other vital ingredient for a massive vector boson is a third polarization state. Similarly, also this other degree of freedom will be provided effectively by the medium, as will be discussed below.

To make the preceding discussion applicable to elementary particle physics, it is useful to write them down covariantly. For this purpose, select the Landau gauge $\partial_\mu A^\mu = 0$. Then the Maxwell equation takes the form

$$\partial^2 A_\mu = j_\mu.$$

The necessary requirement that, at least approximately, the current has a component proportional to A_μ is then directly obtained by considering the current for a charged, scalar field, which has the form

$$j_\mu = iq(\phi^\dagger \partial_\mu \phi - (\partial_\mu \phi^\dagger)\phi) - 2q^2 A_\mu |\phi|^2,$$

which can be derived as the Noether current from the corresponding Lagrangian

$$\begin{aligned} \mathcal{L} &= D_\mu \phi^\dagger D^\mu \phi \\ D_\mu &= \partial_\mu + iqA_\mu. \end{aligned}$$

Thus, the Maxwell equation for A_μ becomes

$$\partial^2 A_\mu = iq(\phi^\dagger \partial_\mu \phi - (\partial_\mu \phi^\dagger)\phi) - 2q^2 A_\mu |\phi|^2.$$

This will have precisely the required form, if the modulus of the field, $|\phi| = v$ can be arranged to have a non-zero value in the vacuum. However, this will only be possible, if its expectation value obeys

$$\langle 0|\phi|0\rangle \neq 0 \tag{6.7}$$

in the selected gauge. In this case, the photon would acquire an effective mass of $(2q^2v^2)^{1/2}$. And by this, the photon is screened out, making the effect short-ranged. How to implement this will be discussed next.

6.5 The Brout-Englert-Higgs effect and the Higgs boson

Motivated by the previous discussion, and by the considerations in section 5.7 it appears reasonable to start with a scalar field with self-interactions, call it Higgs field, and couple it to a gauge field. When coupling the Higgs to a gauge field, symmetry breaking[3], now called the Brout-Englert-Higgs (and several other names) effect, becomes more complicated, but at the same time also more interesting, than in the case without gauge fields.

For simplicity, start with an Abelian gauge theory, coupled to a single, complex scalar, the so-called Abelian Higgs model, before going to the full electroweak and non-Abelian case. This theory has the Lagrangian

$$\begin{aligned}\mathcal{L} &= \frac{1}{2}((\partial_\mu + iqA_\mu)\phi)^\dagger(\partial^\mu + iqA^\mu)\phi - \frac{1}{4}(\partial_\mu A_\nu - \partial_\nu A_\mu)(\partial^\mu A^\nu - \partial^\nu A^\mu)\\ &+\frac{1}{2}\mu^2\phi^\dagger\phi - \frac{1}{2}\frac{\mu^2}{f^2}(\phi^\dagger\phi)^2. \qquad (6.8)\end{aligned}$$

Note that the potential terms are not modified by the presence of the gauge-field. Therefore, the extrema have still the same form and values as in the previous case, at least classically. However, it cannot be excluded that the quartic $\phi^\dagger\phi A_\mu A^\mu$ term strongly distorts the potential. This does not appear to be the case in the electroweak interaction, and this possibility will therefore be ignored.

To make the consequences of the Higgs effect transparent, it is useful to rewrite the scalar field as[4]

$$\phi(x) = \left(\frac{f}{\sqrt{2}} + \rho(x)\right)\exp(i\alpha(x)).$$

This is just another reparametrization for the scalar field, compared to η and ξ previously. It is such that at $\rho = 0$ this field configuration will be a classical minimum of the potential for any value of the phase α. Inserting this parametrization into the Lagrangian (6.8) yields

$$\begin{aligned}\mathcal{L} &= \frac{1}{2}\partial_\mu\rho\partial^\mu\rho + \frac{1}{2}\left(\frac{f}{\sqrt{2}}+\rho\right)^2\partial_\mu\alpha\partial^\mu\alpha + qA^\mu\left(\frac{f}{\sqrt{2}}+\rho\right)^2\partial_\mu\alpha + \frac{q^2}{4}A_\mu A^\mu\left(\frac{f}{\sqrt{2}}+\rho\right)^2\\ &-\frac{1}{4}(\partial_\mu A_\nu - \partial_\nu A_\mu)(\partial^\mu A^\nu - \partial^\nu A^\mu) + \frac{1}{2}\mu^2\left(\frac{f}{\sqrt{2}}+\rho\right)^2 - \frac{1}{2}\frac{\mu^2}{f^2}\left(\frac{f}{\sqrt{2}}+\rho\right)^4\end{aligned}$$

This is an interesting structure, where the interaction pattern of the photon with the radial and angular part are more readily observable.

[3]Actually, symmetry breaking is a misnomer. What happens is more involved, and is explored in the lecture on electroweak physics.

[4]Note that if the space-time manifold is not simply connected and/or contains holes, it becomes important that α is only defined modulo 2π. For flat Minkowski (or Euclidean) space, this is of (almost) no importance. However, it can be important, e. g., in finite temperature calculations. It is definitely important in ordinary quantum mechanics, where, e. g., the Aharanov-Bohm effect and flux quantization depend on this.

Now, it is possible to make the deliberate gauge choice

$$\partial_\mu A^\mu = -\frac{1}{q}\partial^2\alpha. \tag{6.9}$$

This is always possible. It is implemented by first going to Landau gauge and then perform the gauge transformation

$$\begin{aligned} A_\mu &\rightarrow A_\mu + \frac{1}{q}\partial_\mu\alpha \\ \phi &\rightarrow \exp(-i\alpha)\phi. \end{aligned}$$

This gauge choice has two consequences. The first is that it makes the scalar field real everywhere. Therefore, the possibility of selecting the vacuum expectation value of ϕ to be real is a gauge choice. Any other possibilities, e. g. purely imaginary, would be an equally well justified gauge choice. This also implies that the actual value of the vacuum expectation value f of ϕ is a gauge-dependent quantity, and will vanish when not fixing a gauge. The Lagrangian then takes the form[5]

$$\begin{aligned} \mathcal{L} &= \frac{1}{2}\partial_\mu\rho\partial^\mu\rho + \frac{q^2}{4}A_\mu A^\mu\left(\frac{f}{\sqrt{2}}+\rho\right)^2 \\ &-\frac{1}{4}(\partial_\mu A_\nu - \partial_\nu A_\mu)(\partial^\mu A^\nu - \partial^\nu A^\mu) + \frac{1}{2}\mu^2\left(\frac{f}{\sqrt{2}}+\rho\right)^2 - \frac{1}{2}\frac{\mu^2}{f^2}\left(\frac{f}{\sqrt{2}}+\rho\right)^4 \end{aligned} \tag{6.10}$$

i. e., the second term has now exactly the form of a screening term, and yields an effective mass $qf/4$ for the photon field. Furthermore, if ρ could be neglected, the Lagrangian would be just

$$\mathcal{L} = \frac{q^2f^2}{8}A_\mu A^\mu - \frac{1}{4}(\partial_\mu A_\nu - \partial_\nu A_\mu)(\partial^\mu A^\nu - \partial^\nu A^\mu),$$

i. e., the one of a massive gauge field. Together with the gauge condition for A_μ, which links the longitudinal part of the gauge boson to the now explicitly absent degree of freedom α, this implies that the field A_μ acts now indeed as a massive spin-1 field. Furthermore, this Lagrangian is no longer gauge-invariant. This is not a problem, as it was obtained by a gauge choice. Thus, it is said that the gauge symmetry is hidden. Its consequences are still manifest, e. g., the mass for the gauge boson is not a free parameter, but given by the other parameters in the theory. By measuring such relations, it is in principle possible to determine whether a theory has a hidden symmetry or not.

Colloquial, the hiding of a symmetry is also referred to as the breaking of the symmetry in analogy with the case of a global symmetry. However, a theorem, Elitzur's theorem, actually forbids this to be literally true.

[5]Upon quantization, additional terms are introduced due to the gauge-fixing. At the classically level studied here, they can be ignored.

The Goldstone theorem of section 5.7, actually guarantees that a mass will be provided to each gauge boson associated with one of the hidden directions at the classical level. This hidden direction is here the phase, which is massless. This masslessness of the Goldstone boson is actually guaranteed by the Goldstone theorem. Hence, a massless Goldstone boson is effectively providing a mass to a gauge boson by becoming its third component via gauge-rotation, and vanishes by this from the spectrum. It is the tri-linear couplings which provide the explicit mixing terms delivering the additional degree of freedom for the gauge boson at the level of the Lagrangian. Hence, the number of degrees of freedom is preserved in the process: In the beginning there were two scalar and two vector degrees of freedom, now there is just one scalar degree of freedom, but three vector degrees of freedom. The gauge-transformation made nothing more than to shift one of the dynamic degrees of freedom from one field to the other. This was possible due to the fact that both the scalar and the photon are transforming non-trivially under gauge-transformations.

This can, and will be below, generalized for other gauges. In general, it turns out that for a covariant gauge there are indeed six degrees of freedom, four of the vector field, and two from the scalars. Only after calculating a process it will turn out that certain degrees of freedom cancel out, yielding just a system which appears like having a massive vector particle and a single scalar.

Note that though the original scalar field ϕ was charged as a complex field, the radial excitation as the remaining degree of freedom is actually no longer charged: The coupling structure appearing in (6.10) is not the one expected for a charged field.

However, the choice (6.9), which is called the unitary gauge, is extremely intransparent and cumbersome for most actual calculations. The reason is that the quantization of the the theory in this gauge requires the introduction of an infinite number of further terms into the Lagrangian, though with fixed coefficients.

A more convenient possibility, though at the cost of having unphysical degrees of freedom which only cancel at the end, are 't Hooft (or radiative) gauges. To define this gauge once more the decomposition

$$\phi = (\chi, f + \eta)$$

for the scalar field is useful. The Lagrangian then takes the form, up to constant terms,

$$\begin{aligned} \mathcal{L} &= -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{(gf)^2}{2}A_\mu A^\mu + \frac{1}{2}\partial_\mu\eta\partial^\mu\eta - \mu^2\eta^2 + \frac{1}{2}\partial_\mu\chi\partial^\mu\chi - gfA^\mu\partial_\mu\chi \\ &\quad +2igA^\mu(\chi\partial_\mu\chi + \eta\partial_\mu\eta) - 2g(gf)\eta A_\mu A^\mu - g^2(\chi^2+\eta^2)A_\mu A^\mu - \frac{1}{2}\frac{\mu^2}{f^2}(\chi^2+\eta^2)^2. \end{aligned} \tag{6.11}$$

There are a number of interesting observations. Again, there is an effective mass for A_μ, due to the four-field interaction term. Secondly, only the η field has a conventional mass term, the mass-term for the χ field has canceled with the two-field-two-condensate term from the quartic piece of the potential. Finally, there will be terms of type $fA^\mu\partial_\mu\chi$. This implies that a photon can change into a χ while moving, with a strength proportional to the condensate f, and thus mixing between one of the Higgs degrees of freedom and the photon occurs. Therefore, the photon and this scalar, the would-be Goldstone boson, will mix. Note finally that though many more interaction terms have appeared, none of them has any new free parameter, as a consequence of the now hidden symmetry. The only quantity which looks like a new quantity is the condensate value f, though it is classically uniquely determined by the shape of the potential. In a quantum calculation, it can also be determined, but not in perturbation theory, where it remains a fit parameter.

The gauge condition most useful for quantizing this Lagrangian is given by

$$\partial_\mu A^\mu = qf\xi\chi, \tag{6.12}$$

the so-called renormalizable or 't Hooft gauge, where ξ is an arbitrary parameter, a so-called gauge parameter. This gauge choice removes at the quantum level the mixing between the photon and the Higgs. The resulting theory has then the massive Higgs field, a massless Goldstone field, and a massive photon field. The surplus degrees of freedom are, as stated, not actually there, but will cancel at the quantum level.

The form of the Lagrangian (6.11) already indicates how the weak gauge bosons will receive their mass. The necessary Higgs boson has by now indeed been discovered, and its mass is found to be about 125 GeV.

6.6 Parity violation and fermion masses

So far, the construction seems to reproduce very well and elegantly the observed phenomenology of the electroweak interactions.

However, there is one serious flaw. The Dirac equation for fermions, like leptons and quarks, has the form

$$0 = (i\gamma^\mu D_\mu - m)\psi = \left(i\gamma^\mu D_\mu + \frac{1-\gamma_5}{2}m + \frac{1+\gamma_5}{2}m\right)\psi.$$

The covariant derivative is a vector under a weak isospin gauge transformation, and so is the spinor $(1-\gamma_5)/2\psi$. However, the spinor $(1+\gamma_5)/2\psi$ is a singlet under such a gauge transformation. Hence, not all terms in the Dirac equation transform covariantly, and therefore weak isospin cannot be a symmetry for massive fermions. Another way of observing this is that the mass term for fermions in the Lagrangian can be written as

$$\mathcal{L}_{m\psi} = m(\overline{\psi}_L\psi_R - \overline{\psi}_R\psi_L),$$

and therefore cannot transform as a gauge singlet.

However, massless fermions can be accommodated in the theory. Since the observed quarks and (at least almost) all leptons have a mass, it is therefore necessary to find a different mechanism which provides the fermions with a mass without spoiling the isoweak gauge invariance.

A possibility to do so is by invoking the Higgs-effect also for the fermions and not only for the weak gauge bosons. By adding an interaction

$$\mathcal{L}_h = g_f\phi_k\bar{\psi}_i\left(\alpha^k_{ij}\frac{1-\gamma_5}{2}+\beta^k_{ij}\frac{1+\gamma_5}{2}\right)\psi_j+\left(g_f\phi_k\bar{\psi}_i\left(\alpha^k_{ij}\frac{1-\gamma_5}{2}+\beta^k_{ij}\frac{1+\gamma_5}{2}\right)\psi_j\right)^\dagger,$$

this is possible. The constant matrices α and β have to be chosen such that the terms become gauge-invariant. Calculating their precise form is tedious, but straightforward. If, in this interaction, the Higgs field acquires a vacuum expectation value, $\phi = f+$quantum fluctuations, this term becomes an effective mass term for the fermions, and it is trivially gauge-invariant. Alongside with it comes then an interaction of Yukawa-type of the fermions with the Higgs particle. However, the interaction strength is not a free parameter of the theory, since the coupling constants are uniquely related to the tree-level mass m_f of the fermions by

$$g_f = \frac{\sqrt{2}m_f}{f}.$$

But the 12 coupling constants for the three generations of quarks and leptons are not further constrained by the theory, introducing a large number of additional parameters in the theory. Though phenomenologically successful, this is the reason why many feel this type of description of the electroweak sector is insufficient. However, even if it would be incorrect after all, it is an acceptable description at energies accessible so far, and thus has become part of the standard model.

6.7 Weak isospin, hypercharge, and electroweak unification

It is now clear that a theoretical description of the weak interactions requires a gauge theory, as well as some matter field(s) to hide it and provide a mass to the weak gauge bosons. The fact that there are three distinct gauge bosons, two Ws and one Z indicates that the gauge theory has to be more complex than just the U(1) gauge theory of QED. It will thus be a non-Abelian gauge theory. Furthermore, since two of them are charged the connection to the electromagnetic interactions will not be trivial, and there will be some kind of mixing. All of these aspects will be taken into account in the following.

6.7.1 Constructing the gauge group

Phenomenologically, as discussed in section 6.3, the weak interaction provides transitions of two types. One is a charge-changing reaction, which acts between two gauge eigenstates. In case of the leptons, these charge eigenstates are almost (up to a violation of the order of the neutrino masses) exactly a doublet - e. g., the electron and its associated electron neutrino. Therefore, the gauge group of the weak interaction should provide a doublet representation. In case of the quarks this is less obvious, but also they furnish a doublet structure. Hence, an appropriate gauge group for the weak interaction will contain at least SU(2). Since only the left-handed particles are affected, this group is often denoted as $\mathrm{SU}(2)_L$, but this index will be dropped here. Therefore, there are three doublets, generations, of leptons and quarks, respectively in the standard model. However, the mass eigenstates mix all three generations, as will be discussed in detail below.

Since the electric charge of the members of the doublets differ by one unit, the off-diagonal gauge bosons, the W, must carry a charge of one unit. Furthermore, such a gauge group has three generators. The third must therefore be uncharged, as it mediates interactions without exchanging members of a doublet.

The quantum number distinguishing the two eigenstates of a doublet is called the third component of the weak isospin t, and will be denoted by t_3 or I_W^3. Therefore, the gauge group of the weak interactions is called the weak isospin group.

However, the weak gauge bosons are charged. Therefore, ordinary electromagnetic interactions have to be included somehow. Ordinary electromagnetism has as gauge group the Abelian U(1). The natural ansatz for the gauge group of the electroweak interactions is thus the gauge group SU(2)×U(1)[6]. With this second factor-group comes a further quantum number, which is called the hypercharge y. The ordinary electromagnetic charge is then given by

$$eQ = e\left(t_3 + \frac{y}{2}\right). \tag{6.13}$$

Thus, the ordinary electromagnetic interaction must be somehow mediated by a mixture of the neutral weak gauge boson and the gauge boson of the U(1). This is dictated by observation: It is not possible to adjust otherwise the quantum numbers of the particles such that experiments are reproduced. The hypercharge of all left-handed leptons is -1, while the one of left-handed quarks is $y = +1/3$.

Right-handed particles are neutral under the weak interaction. In contrast to the $t = 1/2$ doublets of the left-handed particle, they belong to a singlet, $t = 0$. All in all, the following assignment of quantum numbers for charge, not mass, eigenstates will be

[6] Actually, the correct choice is $\mathrm{SU}(2)/\mathrm{Z}_2\times\mathrm{U}(1)$, although this difference is very rarely of relevance.

necessary to reproduce the experimental findings:

- Left-handed neutrinos: $t = 1/2$, $t_3 = 1/2$, $y = -1$ ($Q = 0$)
- Left-handed leptons: $t = 1/2$, $t_3 = -1/2$, $y = -1$ ($Q = -1$)
- Right-handed neutrinos: $t = 0$, $t_3 = 0$, $y = 0$ ($Q = 0$)
- Right-handed leptons: $t = 0$, $t_3 = 0$, $y = -2$ ($Q = -1$)
- Left-handed up-type (u, c, t) quarks: $t = 1/2$, $t_3 = 1/2$, $y = 1/3$ ($Q = 2/3$)
- Left-handed down-type (d, s, b) quarks: $t = 1/2$, $t_3 = -1/2$, $y = 1/3$ ($Q = -1/3$)
- Right-handed up-type quarks: $t = 0$, $t_3 = 0$, $y = 4/3$ ($Q = 2/3$)
- Right-handed down-type quarks: $t = 0$, $t_3 = 0$, $y = -2/3$ ($Q = -1/3$)
- W^+: $t = 1$, $t_3 = 1$, $y = 0$ ($Q = 1$)
- W^-: $t = 1$, $t_3 = -1$, $y = 0$ ($Q = -1$)
- Z: $t = 1$, $t_3 = 0$, $y = 0$ ($Q = 0$)
- γ: $t = 0$, $t_3 = 0$, $y = 0$ ($Q = 0$)
- Gluon: $t = 0$, $t_3 = 0$, $y = 0$ ($Q = 0$)
- Higgs: a complex doublet, $t = 1/2$ with weak hypercharge $y = 1$. This implies zero charge for the $t_3 = -1/2$ component, and positive charge for the $t_3 = 1/2$ component and negative charge for its complex conjugate

This concludes the list of charge assignments for the standard model particles. The Higgs case will be special, and will be detailed in great length below.

Since at the present time the photon field and the Z boson are not yet readily identified, it is necessary to keep the gauge boson fields for the SU(2) and U(1) group differently, and these will be denoted by W and B respectively. The corresponding pure gauge part of the electroweak Lagrangian will therefore be

$$\begin{aligned}
\mathcal{L}_g &= -\frac{1}{4}G^a_{\mu\nu}G_a^{\mu\nu} - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} \\
F_{\mu\nu} &= \partial_\mu B_\nu - \partial_\nu B_\mu \\
G^a_{\mu\nu} &= \partial_\mu W^a_\nu - \partial_\nu W^a_\mu + g f^{abc} W^b_\mu W^c_\nu,
\end{aligned}$$

where g is the weak isospin gauge coupling. f^{abc} are the structure constants of the weak isospin gauge group, which is just the SU(2) gauge group.

Coupling matter fields to these gauge fields proceeds using the ordinary covariant derivative, which takes the form

$$D_\mu = \partial_\mu + \frac{ig}{2}\tau_a W_\mu^a + \frac{ig'y}{2}B_\mu,$$

where g' is the hypercharge coupling constant, which is modified by the empirical factor y. Note that y is not constrained by the gauge symmetries, and its value is purely determined by experiment. Thus, why it takes the rational values it has is an unresolved question to date. However, this fact would come about naturally, if the weak gauge group would originate from a different gauge group at higher energies, say SU(2)×U(1)⊂SU(3), which is hidden to the extent that all other fields charged under this larger gauge group are effectively so heavy that they cannot be observed with current experiments.

The matrices τ_a are determined by the representation of SU(2) in which the matter fields are in. For a doublet, these will be the Pauli-matrices. For the adjoint representation, these would be given by the structure constants, $\tau^a_{bc} = f^{abc}$, and so on. For fermions, of course, this covariant derivative is contracted with the Dirac matrices γ_μ. Precisely, to couple only to the left-handed spinors, it will be contracted with $\gamma_\mu(1-\gamma_5)/2$ for the W_μ^a term and with γ_μ for the kinetic and hypercharge term. By this, the phenomenological couplings are recovered in the low-energy limit, as the exchange of a massive gauge boson then becomes proportional to $1/M^2$, thus recovering the Fermi-coupling g^2/M^2.

6.7.2 Hiding the electroweak symmetry

To have a viable theory of the electroweak sector it is necessary to hide the symmetry such that three gauge bosons become massive, and one uncharged one remains massless. Though this can be of course arranged in any gauge, it is most simple to perform this in the unitary gauge. Since three fields have to be massive, this will require three pseudo-Goldstone bosons. Also, since empirically two of them have to be charged, as the $W^\pm$ bosons are charged, the simplest realization is by coupling a complex doublet scalar field, the Higgs field, to the electroweak gauge theory

$$\begin{aligned}
\mathcal{L}_h &= \mathcal{L}_g + (D_\mu\phi)^\dagger D^\mu\phi + V(\phi\phi^\dagger) \\
\phi &= \frac{1}{\sqrt{2}}\begin{pmatrix}\phi_1 + i\phi_2 \\ \phi_3 + i\phi_4\end{pmatrix}
\end{aligned} \tag{6.14}$$

where the field ϕ is thus in the fundamental representation of the gauge group. Its hypercharge will be 1, an assignment which will be necessary below to obtain a massless photon. The potential V can only depend on the gauge-invariant combination $\phi^\dagger\phi$, and thus can, to be renormalizable, only contain a mass-term and a quartic self-interaction. The mass-term must be again of the wrong sign (imaginary mass), such that there exists a possibility for the ϕ field to acquire a (gauge-dependent) vacuum-expectation value.

To work in unitary gauge it is best to rewrite the Higgs field in the form

$$\phi = e^{\frac{i\tau^a\alpha_a}{2}}\begin{pmatrix}0\\ \rho\end{pmatrix}.$$

There are now the three α^a fields and the ρ field. Performing a gauge transformation such that the phase becomes exactly canceled, and setting $\rho = f + \eta$ with f constant makes the situation similar to the one in the Abelian Higgs model. Note that by a global gauge transformation the component with non-vanishing expectation value can be selected still at will.

The mass will be made evident by investigating the equation of motion for the gauge bosons

$$\begin{aligned}\partial^2 W_\mu^a - \partial_\mu(\partial^\nu W_\nu^a) &= j_\mu^{a\phi} + j_\mu^{aW}\\ \partial^2 B_\mu - \partial_\mu(\partial^\nu B_\nu) &= j_\mu^y,\end{aligned}$$

where j^W contains contributions which involve only the Ws, while j^ϕ contains all terms from the Higgs, and contributions from the fermions are neglected for now. Only the Higgs contributions will be relevant below, and therefore for it the subscript ϕ will be dropped. Thus, it is necessary to determine the weak isospin current j_μ^a and the hypercharge current j_μ^y. In contrast to the Abelian case this current will now also contain contributions from the self-interactions of the W gauge bosons. However, at tree-level these can be ignored.

The contribution of the Higgs field to the currents[7] are

$$\begin{aligned}j_\mu^a &= \frac{ig}{2}(\phi^\dagger\tau^a D_\mu\phi - (D_\mu\phi)^\dagger\tau^a\phi)\\ j_\mu^y &= \frac{ig'}{2}(\phi^\dagger D_\mu\phi - (D_\mu\phi)^\dagger\phi).\end{aligned}$$

The appearance of τ^a makes manifest that the current j_μ^a is a weak isovector current, while their absence signifies that the hypercharge current j_μ^y is a weak isoscalar current.

Expanding the covariant derivatives yields[8]

$$\begin{aligned}j_\mu^a &= \frac{ig}{2}(\phi^\dagger\tau^a\partial_\mu\phi - (\partial_\mu\phi)^\dagger\tau^a\phi) - \frac{g^2}{2}\phi^\dagger\phi W_\mu^a - \frac{gg'y}{2}\phi^\dagger\tau^a\phi B_\mu \qquad (6.15)\\ j_\mu^y &= \frac{ig'}{2}(\phi^\dagger\partial_\mu\phi - (\partial_\mu\phi)^\dagger\phi) - \frac{gg'y}{2}\phi^\dagger\tau_a W_\mu^a\phi - \frac{(g'y)^2}{2}\phi^\dagger\phi B_\mu.\end{aligned}$$

[7] Of course, all fields with non-zero weak isospin will contribute to this current. But only the Higgs-field will do so classically.

[8] Using $\{\tau^a,\tau^b\} = 2\delta^{ab}$.

It should be noted that the two equations are not only coupled by the Higgs field, but by both, the W_μ^a and B_μ gauge fields. Selecting now the Higgs field to be of the form

$$\phi = \begin{pmatrix} 0 \\ \frac{f}{\sqrt{2}} \end{pmatrix} + \text{quantum fluctuations},$$

where the quantum fluctuations vanish at tree-level will provide the vacuum expectation value

$$\langle 0|\phi|0\rangle = \begin{pmatrix} 0 \\ \frac{f}{\sqrt{2}} \end{pmatrix}.$$

This is then as in the Abelian case, and indeed sufficient to provide a mass for the gauge bosons.

Ignore for a moment the hypercharge contribution. Then the term linear in W_μ^a on the right-hand side of (6.15) provides a term

$$-M^2 W_\mu^a = -\left(\frac{gf}{2}\right)^2 W_\mu^a,$$

and thus a mass $M_W = gf/2$ is provided. Thus, the equation of motion for the W_μ^a field takes the form

$$(\partial^2 + M_W^2)W_\mu^a - \partial_\mu(\partial^\nu W_\nu^a) = j_\mu^a(f, \phi, W_\mu^a, ...),$$

where the mass-term has been removed from the current. Thus, formally this is the equation of motion for a massive gauge boson, interacting with itself and other fields due to the current j_μ^a. As has been seen in the Abelian case, the additional degree of freedom has been selected by the choice of gauge, and is provided by the phase of the Higgs field.

Next it is necessary to check what happens when including also the B_μ part once more. Since the $t_3 = -1/2$ component of the Higgs field delivers the mass, this component should not be electrically charged. According to the relation (6.13) thus the assignment of the hypercharge $y = 1$ for the Higgs field is a-posterior justified. In the full current of the Higgs field, besides the contribution $\phi^\dagger\phi W_\mu^a$ also the contribution $gg'y/2B_\mu\phi^\dagger\tau^a\phi$ appears. Since the matrices τ^1 and τ^2 are off-diagonal it follows that

$$(0, f)\tau^{1,2}\begin{pmatrix} 0 \\ f \end{pmatrix} = 0.$$

Thus only in the third component of the current a contribution due to the vacuum expectation value of the Higgs field appears. As a consequence, despite the appearance of the additional B_μ gauge boson the result for the mass (at tree-level) for the off-diagonal

$W_\mu^{1,2}$ bosons is not changed. However, this is not the case for the third component. The relevant part of the equation of motion reads

$$\partial^2 W_\mu^3 - \partial_\mu(\partial^\nu W_\nu^3) = -\left(\frac{g^2 f^2}{2}\right)^2 W_\mu^3 + \frac{gf}{2}\frac{g'f}{2}B_\mu.$$

Consequently, also the relevant part of the equation of motion for the hypercharge gauge boson is of similar structure

$$\partial^2 B_\mu - \partial_\mu(\partial^\nu B_\nu) = -\left(\frac{g'f}{2}\right)^2 B_\mu + \frac{gf}{2}\frac{g'f}{2}W_\mu^3.$$

Hence, even in the vacuum the equations are coupled, and thus both gauge bosons mix. These equations can be decoupled by changing variables as

$$\begin{aligned} A_\mu &= W_\mu^3 \sin\theta_W + B_\mu \cos\theta_W \\ Z_\mu &= W_\mu^3 \cos\theta_W - B_\mu \sin\theta_W, \end{aligned}$$

which are the fields given the name of the photon A_μ and the Z boson Z_μ. The mixing parameter θ_W is the (Glashow-)Weinberg angle θ_W, and is given entirely in terms of the coupling constants g and g' as

$$\begin{aligned} \tan\theta_W &= \frac{g'}{g} \\ \cos\theta_W &= \frac{g}{\sqrt{g^2+g'^2}} \\ \sin\theta_W &= \frac{g'}{\sqrt{g^2+g'^2}}. \end{aligned}$$

Using the inverse transformations

$$\begin{aligned} W_\mu^3 &= A_\mu \sin\theta_W + Z_\mu \cos\theta_W \\ B_\mu &= A_\mu \cos\theta_W - Z_\mu \sin\theta_W \end{aligned}$$

it is possible to recast the equations of motion. The equation of motion for the photon takes the form

$$\partial^2 A_\mu - \partial_\mu(\partial^\nu A_\nu) = \sin\theta_W j_\mu^3.$$

Here, j_μ^3 contains all contributions which are not providing a mass to the gauge bosons. Thus, the photon is effectively massless. The price paid is that it now has also direct interactions with the weak gauge bosons. On the other hand, the equation for the Z boson takes the form

$$\partial^2 Z_\mu - \partial_\mu\partial^\nu Z_\nu = -\frac{f^2(g^2+g'^2)}{4}Z_\mu + \cos\theta_W j_\mu^3.$$

Thus, the Z boson acquires a mass larger than the $W^{1,2}$ boson. In particular the relation is

$$M_Z \cos\theta_W = M_W.$$

According to the best measurements this effect is, however, only of order 15%. At the same time the self-interaction of the Z boson with the other weak bosons is reduced compare to the one of the original W_μ^3 boson.

Of course, this changes also the form of the coupling to the neutral fields for matter. In particular, the symmetry of SU(2) is no longer manifest, and the $W^{1,2}$ cannot be treated on the same footing as the neutral bosons. E. g., the neutral part of the coupling in the covariant derivative now takes the form

$$D_\mu^{(N)} = \partial_\mu + ig\sin\theta_W A_\mu \left(t_3 + \frac{y}{2}\right) + i\frac{g}{\cos\theta_W} Z_\mu \left(t_3\cos^2\theta_W - \frac{y}{2}\sin^2\theta_W\right).$$

Using the relation (6.13), it is possible to identify the conventional electric charge as

$$e = g\sin\theta_W, \tag{6.16}$$

i. e., the observed electric charge is smaller than the hypercharge. It should be noted that this also modifies the character of the interaction. While the interaction with the photon is purely vectorial, and the one with the $W^{1,2}$ bosons remains left-handed (axial-vector), the interaction with the Z boson is now a mixture of both, and the mixing is parametrized by the Weinberg angle.

Due to the mixing the original 3-W and 4-W interactions now also involve the photon field A_μ. Redefining

$$W^\pm = \frac{1}{\sqrt{2}}(W^1 \pm iW^2)$$

shows that the couplings between the new fields $W^\pm$ and the photon A_μ are such as if the $W^\pm$ carry a charge of one in units of the electron charge (6.16). Thus, they can be regarded as electromagnetically charged fields.

Note that the masslessness of the photon is directly related to the fact that the corresponding component of the Higgs-field has no vacuum expectation value,

$$\left(\frac{y}{2} + \frac{\tau^3}{2}\right)\langle 0|\phi|0\rangle = 0,$$

and thus the vacuum is invariant under a gauge transformation involving a gauge transformation of A_μ,

$$\langle 0|\phi'|0\rangle = \langle 0|\exp(i\alpha(y/2+\tau_3))\phi|0\rangle.$$

Thus, the original SU(2)×U(1) gauge group is hidden, and only a particular combination of the subgroup U(1) of SU(2) and the factor U(1) is not hidden, but a manifest gauge symmetry of the system, and thus this U(1) subgroup is the stability group of the electroweak gauge group SU(2)×U(1). It is said, by an abuse of language, that the gauge group SU(2)×U(1) has been broken down to U(1). Since this gauge symmetry is manifest, the associated gauge boson, the photon A_μ, must be massless. If, instead, one would calculate without the change of basis, none of the gauge symmetries would be manifest. However, the mixing of the B_μ and the W_μ^3 would ensure that at the end of the calculation everything would come out as expected from a manifest electromagnetic symmetry.

Also, this analysis is specific to the unitary gauge. In other gauges the situation may be significantly different formally. Only when determining gauge-invariant observables, like scattering cross-sections or the masses of gauge invariant bound-states, like positronium, everything will be the same once more.

6.8 CP violation and the CKM matrix

In any strong or electromagnetic process the quark (and lepton) flavor is conserved. E. g., the strangeness content is not changing. This is not true for weak processes. It is found that they violate the flavor number of both, quarks and leptons. In case of the leptons, this effect is suppressed by the small neutrinos masses involved, but in case of quarks this is a significant effect.

Considering the weak decays of a neutron compared to that of a strange Λ, it is found that the relative strengths can be expressed.as

$$\begin{aligned} g_{\Delta S=0} &= g' \cos\theta_C \\ g_{\Delta S=1} &= g' \sin\theta_C \end{aligned}$$

where g' is a universal strength parameter for the weak interactions, its coupling constant. The angle parameterizing the decay is called the Cabibbo angle. A similar relation also holds in the leptonic sector for the muon quantum number

$$\begin{aligned} g_{\Delta\mu=0} &= g' \cos\theta_C^L \\ g_{\Delta\mu=1} &= g' \sin\theta_C^L, \end{aligned}$$

where, however, $\sin\theta_C^L$ is almost one, while in the quark sector $\sin\theta_C$ is about 0.22. Corresponding observations are also made for other flavors. This also explains the intergeneration violation of flavor, the so-called Glashow-Iliopoulos-Maiani (GIM) mechanism, which incidentally predicted the charm quarks.

This result implies that the mass eigenstates of the matter particles are not at the same time also weak eigenstates, but they mix. Hence, on top of the P-violating and

C-violating factors of $(1-\gamma_5)/2$, it is necessary to include something into the interaction which provides this mixing. This can be done by introducing a flavor-dependent unitary coupling matrix

$$G' = \frac{g'}{\sqrt{2}} \begin{pmatrix} 0 & \cos\theta_C^{(L)} & 0 \\ \cos\theta_C^{(L)} & 0 & \sin\theta_C^{(L)} \\ 0 & -\sin\theta_C^{(L)} & 0 \end{pmatrix}.$$

This is equivalent to just use a doublet

$$\begin{pmatrix} u & d\cos\theta_C + s\sin\theta_C \end{pmatrix},$$

e. g., in the quark sector. Such a doublet structure can be associated with (weak) charges $Q_u = 1/2$ and $Q_{ds} = -1/2$. This is just the weak isospin. It should be noted that the up-type and down-type flavors of each generation are already mixed by the weak interactions themselves, since up-type and down-type are just the two different charge states of the weak isospin. In contrast to this intrageneration mixing, the effect here mixes different generations, and is hence an intergeneration effect, apart from the weak interaction.

Hence, the flavor (and mass) eigenstates of the quarks are effectively rotated by a unitary matrix. For two generations, this matrix, the Cabibbo matrix, is uniquely given by

$$V_C = \begin{pmatrix} \cos\theta_C & \sin\theta_C \\ -\sin\theta_C & \cos\theta_C \end{pmatrix},$$

with again the Cabibbo angle θ_C, with a value of about $\sin\theta_C \approx 0.22$. For three generations, there exist no unique parametrization of the mixing matrix, which is called the Cabibbo-Kobayashi-Maskawa (CKM) matrix. The interaction is thus parametrized as

$$\begin{aligned} &(\bar{u}_L, \bar{c}_L, \bar{t}_L)^T \gamma^\mu W_\mu^\dagger V_{\text{CKM}} \begin{pmatrix} d_L \\ s_L \\ b_L \end{pmatrix} + \text{h.c.} \\ V_{\text{CKM}} &= V_L^u V_L^{d\dagger} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix}. \end{aligned}$$

The absolute values of the entries have been measured independently, and are given (2015 values)2 by

$$\begin{array}{ccc} |V_{ud}| = 0.9743(3) & |V_{us}| = 0.2252(9) & |V_{ub}| = 4.2(5)\times 10^{-3} \\ |V_{cd}| = 0.23(2) & |V_{cs}| = 1.01(3) & |V_{cb}| = 41(1)\times 10^{-3}\,. \\ |V_{td}| = 8.4(6)\times 10^{-3} & |V_{ts}| = 43(3)\times 10^{-3} & |V_{tb}| = 0.89(7) \end{array}$$

Thus, the CKM matrix is strongly diagonal-dominant, especially towards larger quark masses, and within errors compatible with a unitary matrix.

To make the unitarity more explicit, it is common to recast the CKM matrix into the following form

$$\begin{aligned} V_{CKM} &= \begin{pmatrix} c_{12}c_{13} & s_{12}c_{13} & e^{-i\delta_{13}}s_{13} \\ -s_{12}c_{23}-e^{i\delta_{13}}c_{12}s_{23}s_{13} & c_{12}c_{23}-e^{i\delta_{13}}s_{12}s_{23}s_{13} & s_{23}c_{13} \\ s_{12}s_{23}-e^{i\delta_{13}}c_{12}c_{23}s_{13} & -c_{12}s_{23}-e^{i\delta_{13}}s_{12}c_{23}s_{13} & c_{23}c_{13} \end{pmatrix}. \quad (6.17) \\ c_{ij} &= \cos\theta_{ij} \\ s_{ij} &= \sin\theta_{ij} \end{aligned}$$

To have only 4 free parameters (θ_{12}, θ_{13}, θ_{23}, and δ_{13}) requires not only unitarity, but also to exploit some freedom in redefining the absolute phases of the quark fields. Testing whether this matrix is indeed unitary by measuring the nine components individual is currently recognized as a rather sensitive test for physics beyond the standard model. The condition for unitarity can be recast into a form that a sum of three functions of the angles has to be π, forming a triangle. Thus, testing for the unitarity, and thus standard-model compatibility of CP violations, is often termed measuring the unitarity triangle.

The presence of this matrix also gives rise to the possibility that not only C and P are violated separately, but that also the compound symmetry CP is violated (and therefore also T by virtue of the CPT-theorem). That occurs to lowest order in perturbation theory by a box-diagram exchanging the quark flavor of two quarks by the exchange of two W bosons.

That such a process violates CP can be seen as follows. The process just described is equivalent to the oscillation of, e. g., a $d\bar{s}$ bound state into a $s\bar{d}$ bound state, i. e., a neutral kaon K^0 into its anti-particle $\bar{K}^0$. The C and P quantum numbers of both particles are P$=-1$, C$=1$ and P$=1$, C$=1$, respectively, and thus CP$=-1$ and CP$=1$. Thus, any such transition violates CP. Performing the calculation of the corresponding diagram yields that it is proportional to the quantity

$$\begin{aligned} \chi &= \sin\theta_{12}\sin\theta_{23}\sin\theta_{13}\cos\theta_{12}\cos\theta_{23}\cos^2\theta_{13}\sin\delta_{13} \quad (6.18) \\ &\quad \times(m_t^2-m_c^2)(m_t^2-m_u^2)(m_c^2-m_u^2)(m_b^2-m_s^2)(m_b^2-m_d^2)(m_s^2-m_d^2). \end{aligned}$$

Thus, such a process, and thus CP violation is possible if there is mixing at all (all θ_{ij} non-zero) with a non-trivial phase δ_{13}, and the masses of the quarks with fixed charge are all different. They may be degenerate with ones of different charge, however. Since such oscillations are experimentally observed, this already implies the existence of a non-trivial quark-mixing matrix. The value of δ_{13} is hard to determine, but is roughly 1.2.

A rough estimate can be obtained in the following way. The neutral kaon K_0 and its anti-particle $\bar{K}_0$ have, as noted, in the naive quark picture the quark content $d\bar{s}$ and $s\bar{d}$, respectively. Since strangeness is no longer conserved, these states can mix, as $2^{-\frac{1}{2}}(|K_0\rangle \pm |\bar{K}_0\rangle)$. One of this state will decay quickly into two pions in 10^{-11} and is therefore called K_S, and the other in 10^{-8} seconds in three pions or a pair of leptons, called K_L. The reason is that the mixtures are approximate eigenstates of CP, with values +1 and -1. Since the final state has likewise definite CP, and CP is approximately conserved, these are the decay patterns. The time difference for the K_L then comes from the smaller amount of phase space or the smaller electromagnetic force.

CP violation is now seen as follows in this system. The kaons decay, and the kaon amplitude as a function of time will decay over time, which can be modeled as

$$a_i(t) = a_i(0)\exp(-(\Gamma_i/2 - im_i)t),$$

where m is the mass and Γ the inverse decay time, and i enumerates the type. It is possible to create a beam of pure K_0 as the initial state. Because of CP violation, the two states can oscillate into each other, yielding a final intensity after a time t of

$$I = (a_1(t) - a_2(t))(a_1(t) - a_2(t))^* = \frac{1}{4}\left(e^{-\Gamma_1 t} + e^{-\Gamma_2 t} - 2e^{\frac{\Gamma_1+\Gamma_2}{2}t}\cos\Delta mt\right),$$

which is parametrized by the mass difference of both states Δm. This quantity can be directly brought into a relationship with (6.18), and therefore permits a measurement of δ_{13}. The actual value of Δm is tiny, 3.5×10^{-6} eV.

Note that because of the violation of CP, also K_L can decay directly into two pions. However, this effect is strongly suppressed in the standard model.

Within the standard model, there is no explanation of this mixing, however, and thus these are only parameters.

6.9 Neutrino oscillations and the PMNS matrix

A consequence of the mixing by the CKM matrix is that it is possible to start at some point with a particle of any flavor, and after a while it has transformed into a particle of a different flavor. The reason for this is a quantum effect, and proceeds essentially through emission and reabsorption of a virtual $W^\pm$. In quantum mechanical terms for a two-body system, this can be easily deduced. Take a Hamiltonian H as

$$H = \begin{pmatrix} H_0 & \Delta \\ \Delta & H_0 \end{pmatrix}.$$

This leads to a time-evolution operator

$$U(t) = e^{-iH_0 t} \begin{pmatrix} \cos(\Delta t) & -i\sin(\Delta t) \\ -i\sin(\Delta t) & \cos(\Delta t) \end{pmatrix}.$$

Under time evolution an initial pure state $(1, 0)$ will therefore acquire a lower component, if Δ is non-zero. On the other hand, if the composition of a pure state after an elapsed time is measured, it is possible to obtain the size of Δ. The probability P to find a particle in the state $(0, 1)$ after a time t is given by

$$P = \sin^2(\Delta t). \tag{6.19}$$

In the standard model, the corresponding expression for the transition probability involves the mass difference between the two states. To lowest order it is given for the two-flavor case by

$$P = \sin^2\left(\frac{\Delta_m^2 L}{4E}\right)\sin^2(2\theta),$$

where Δ_m is the mass difference between both states, E is the energy of the original particle, L is the distance traveled, and θ is the corresponding Cabibbo angle. If the probability, the energy, and the distance is known for several cases, both the Cabibbo angle and the mass difference can be obtained. Of course, both states have to have the same conserved quantum numbers, like electrical charge.

The same calculation can be performed in the more relevant three-flavor case. The obtained transition probability from a flavor f_i to a flavor f_j is more lengthy, and given by

$$P(f_i \to f_j, L, E) = \sum_k |V_{jk}|^2 |V_{ik}|^2 + 2\sum_{k>l} |V_{jk}V_{ik}^* V_{il} V_{jl}^*| \cos\left(\frac{\Delta_{m_{kl}}^2}{2E}L - \arg(V_{jk}V_{ik}^* V_{il} V_{jl}^*)\right)$$

This sum shows that the process is sensitive to all matrix elements, including the CP-violating phase, of the CKM matrix, while it cannot determine the sign of the mass differences.

For the quarks, the first observation of CKM effects were due to decays. Such oscillations have also been found later, and studied to determine the matrix elements more precisely. E. g., as discussed in section 6.8 it is possible with a certain probability that a particle oscillates from a short-lived state to a long-lived state. This is the case for the so-called K-short K_S and K-long K_L kaons, mixed bound states of an (anti-)strange and an (anti-)down quark. The oscillations are a different way to measure the mass difference. This has been experimentally observed, but the distance L is of laboratory size, about 15 m for K_L and 15 cm for K_S, giving again a Δ_m of about 3.5×10^{-12} MeV. However, in this case the effect is rather used for a precision determination of the mixing angle, since the mass can be accurately determined using other means.

In the lepton sector, these oscillations were the original hint for non-zero neutrino masses. As it is visible, such oscillations only occur if the involved particles have a mass. Thus, only with massive neutrinos there was a possibility to see the effect. But there is no direct determination of their mass, and the best result so far is an upper limit on the order of 2 eV from the β-decay of tritium. Observing these oscillations then indicated the presence of neutrino masses, and that there is a CKM-like matrix also in the lepton sector, as also otherwise there would be no mixing - for a unit matrix the above given formula reduces to the first term only.

With this, it is possible to determine the mass difference of neutrinos. It is found that $|\Delta_{m_{12}}| = 0.0087$ eV and $|\Delta_{m_{23}}| = 0.049$ eV. As only the squares can be determined, it is so far not possible to establish which neutrino is the heaviest, and if one of them is massless. Still, the mass difference of 0.05 eV indicates that with an increase in sensitivity by a factor of 40 it can be determined in decay experiments whether the electron neutrino is the heaviest. If it is, the mass hierarchy is called inverse, as the first flavor, corresponding to the electron, is heaviest. Otherwise, it would be called a normal hierarchy. As a side remark, these tiny mass differences imply that the oscillation lengths L are typically macroscopically, and of the order of several hundred kilometers and more for one oscillation.

Aside from these mass differences, it was also possible to determine some of the elements of the CKM matrix in the lepton sector, which is called Pontecorvo-Maki-Nakagawa-Sakata, or PMNS, matrix. In terms of the parametrization (6.17), the values of the known three angels (as of 2017) are

$$\begin{aligned}\sin^2\theta_{12} &= 0.31(2)\\ \sin^2\theta_{23} &= 0.4(1)\\ \sin^2\theta_{13} &= 0.025(4)\end{aligned}$$

while the value of δ_{13} was not yet measured, and it is thus not clear whether there is also CP violation in the lepton sector. Though the data is slightly in favor of it. The values of the other angles imply that with the sensitivity of current experiments, it should be possible to get a reasonable estimate of δ_{13} until the mid 2020s.

The values of the three real angles are quite astonishing. They imply that, irrespective of the value of δ_{13}, the PMNS matrix is not strongly diagonal dominant. Thus, the lepton flavors mix much stronger than the corresponding quark flavors, and only the very small values of the neutrino masses reduce the oscillation probabilities so strongly that the effect is essentially not seen for electrons, muons, and tauons, and even requires the very weakly interacting neutrinos, which can move over long distances, to observe it at all.

Again, at the current time there is no understanding for why this is the case, nor why the quark sector and the lepton sector are so extremely different also in this respect, and not only for the hierarchy of masses.

6.10 The weak interactions as a gauge theory

Combining everything, the electroweak sector of the standard model, the Glashow-Salam-Weinberg theory is of the following type: It is based on a gauge theory with a gauge group G, which is SU(2)×U(1), weak isospin and hypercharge. There are left-handed and right-handed fermions included, which are differently charged under the gauge group G. These are doublets for the left-handed quarks and leptons, and singlets for the right-handed quarks and leptons[9] for the weak isospin. There is also the scalar Higgs field which is again a doublet.

To formulate the Lagrangian of the electroweak sector of the standard model requires hence a number of steps. First of all, the vacuum expectation value for the Higgs field is chosen to be $f/\sqrt{2}$. Its direction is chosen such that it is manifest electrically neutral. That is provided by the requirement

$$Q\phi = 0 = \left(\frac{\tau_3}{2} + \frac{y_\phi}{2}\right)\phi = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} = 0.$$

The field is then split as

$$\phi = \begin{pmatrix} \phi^+ \\ \frac{1}{\sqrt{2}}(f + \eta + i\chi) \end{pmatrix},$$

with $\phi^- = (\phi^+)^\dagger$, i. e., its hermitian conjugate. The fields ϕ^+ and ϕ^- carry integer electric charge plus and minus, while the fields η and χ are neutral. Since a non-vanishing value of f leaves only a U(1) symmetry manifest, $\phi^\pm$ and χ are the would-be Goldstone bosons. This leads to the properties of the Higgs fields and vector bosons as discussed previously.

The fermions appear as left-handed doublets in three generations

$$\begin{aligned} L_i^L &= \begin{pmatrix} \nu_i^L \\ l_i^L \end{pmatrix} \\ Q_i^L &= \begin{pmatrix} u_i^L \\ d_i^L \end{pmatrix}, \end{aligned}$$

[9]Right-handed neutrinos are a special issue, to be discussed in more detail in section 8.5.4. They will be assumed to exist in this chapter.

where i counts the generations, l are the leptons e, μ and τ, ν the corresponding neutrinos ν_e, ν_μ, ν_τ, u the up-type quarks u, c, t, and d the down-type quarks d, s, b. Correspondingly exist the right-handed singlet fields l_i^R, ν_i^R, u_i^R, and d_i^R. Using this basis the Yukawa interaction part reads

$$\mathcal{L}_Y = -\bar{L}_i^L G_{ij}^{lr} l_j^R \phi^r + \bar{L}_i^L G_{ij}^{\nu r} \nu_j^R \phi^r - \bar{Q}_i^L G_{ij}^{ur} u_j^R \phi^r + \bar{Q}_i^L G_{ij}^{dr} d_j^R \phi^r + h.c..$$

The matrices G are connected to mass matrices by

$$M_{ij}^l = \frac{1}{\sqrt{2}} G_{ij}^l f \qquad M_{ij}^\nu = \frac{1}{\sqrt{2}} G_{ij}^\nu f \qquad M_{ij}^u = \frac{1}{\sqrt{2}} G_{ij}^u f \qquad M_{ij}^d = \frac{1}{\sqrt{2}} G_{ij}^d f.$$

It is then possible to transform the fermion fields[10] f into eigenstates of these mass-matrices, and thus mass eigenstates, by a unitary transformation

$$\begin{aligned} f_i^{fL} &= U_{ik}^{fL} f_k^L \\ f_i^{fR} &= U_{ik}^{fR} f_k^R, \end{aligned} \tag{6.20}$$

for left-handed and right-handed fermions respectively, and f numbers the fermion species l, ν, u, and d and i the generation. The fermion masses are therefore

$$m_{fi} = \frac{1}{\sqrt{2}} \sum_{km} U_{ik}^{fL} G_{km}^f (U^{fR})_{mi}^\dagger f.$$

In this basis the fermions are no longer charge eigenstates of the weak interaction, and thus the matrices U correspond to the CKM matrices. In fact, in neutral interactions which are not changing flavors always combinations of type $U^{fL}(U^{fL})^\dagger$ appear, and thus they are not affected. For flavor-changing (non-neutral) currents the matrices

$$\begin{aligned} V^q &= U^{uL}(U^{dL})^\dagger \\ V^l &= U^{\nu L}(U^{lL})^\dagger \end{aligned} \tag{6.21}$$

remain, providing the flavor mixing. Finally, the electric charge is given by

$$e = \sqrt{4\pi\alpha} = g' \sin\theta_W = g \cos\theta_W,$$

with the standard value $\alpha \approx 1/137$.

[10]The vacuum expectation value f of the Higgs field, carrying no indices, should not be confused with the fermion fields f_i^f, which carry various indices, f denoting the fermion class.

Putting everything together, the lengthy Lagrangian for the electroweak standard model emerges:

$$
\begin{aligned}
\mathcal{L} &= \bar{f}_i^{rs}(i\gamma^\mu\partial_\mu - m_f)f_i - eQ_r\bar{f}_i^{rs}\gamma^\mu f_i^{rs}A_\mu \\
&+ \frac{e}{\sin\theta_W\cos\theta_W}(I^3_{Wr}\bar{f}_i^{rL}\gamma^\mu f_i^{rL} - \sin^2\theta_W Q_r\bar{f}_i^{rs}\gamma^\mu f_i^{rs})Z_\mu \\
&+ \frac{e}{\sqrt{2}\sin\theta_W}(\bar{f}_i^{rL}\gamma^\mu V^r_{ij}f_j^{rL}W^+_\mu + \bar{f}_i^{rL}\gamma^\mu(V^r)^+_{ij}f_j^{rL}W^-_\mu) \\
&- \frac{1}{4}|\partial_\mu A_\nu - \partial_\nu A_\mu - ie(W^-_\mu W^+_\nu - W^-_\nu W^+_\mu)|^2 \\
&- \frac{1}{4}\left|\partial_\mu Z_\nu - \partial_\nu Z_\mu + ie\frac{\cos\theta_W}{\sin\theta_W}(W^-_\mu W^+_\nu - W^-_\nu W^+_\mu)\right|^2 \\
&- \frac{1}{2}\left|\partial_\mu W^+_\nu - \partial_\nu W^+_\mu - ie(W^+_\mu A_\nu - W^+_\nu A_\mu) + ie\frac{\cos\theta_W}{\sin\theta_W}(W^+_\mu Z_\nu - W^+_\nu Z_\mu)\right|^2 \\
&+ \frac{1}{2}\left|\partial_\mu(\eta + i\chi) - i\frac{e}{\sin\theta_W}W^-_\mu\phi^+ + iM_Z Z_\mu + \frac{ie}{2\cos\theta_W\sin\theta_W}Z_\mu(\eta + i\chi)\right|^2 \\
&+ |\partial_\mu\phi^+ + ieA_\mu\phi^+ - ie\frac{\cos^2\theta_W - \sin^2\theta_W}{2\cos\theta_W\sin\theta_W}Z_\mu\phi^+ - iM_W W^+_\mu - \frac{ie}{2\sin\theta_W}W^+_\mu(\eta + i\chi)|^2 \\
&- f^2\eta^2 - \frac{ef^2}{\sin\theta_W M_W}\eta(\phi^-\phi^+ + \frac{1}{2}|\eta + i\chi|^2) \\
&- \frac{e^2 f^2}{4\sin^2\theta_W M_W^2}(\phi^-\phi^+ + \frac{1}{2}|\eta + i\chi|^2)^2 \\
&- \frac{em_{ri}}{2\sin\theta_W M_W}(\bar{f}_i^{rs}f_i^{rs}\eta - 2I^3_{Wr}i\bar{f}_i^{rs}\gamma_5 f_i^{rs}\chi) \\
&+ \frac{e}{\sqrt{2}\sin\theta_W}\frac{m_{ri}}{M_W}(\bar{f}_i^{rR}V^r_{ij}f_j^{rL}\phi^+ + \bar{f}_i^{rL}(V^r)^+_{ij}f_j^{rR}\phi^-) \\
&+ \frac{e}{\sqrt{2}\sin\theta_W}\frac{m_{ri}}{M_W}(\bar{f}_i^{rL}V^r_{ij}f_j^{rR}\phi^+ + \bar{f}_i^{rR}(V^r)^+_{ij}f_j^{rL}\phi^-),
\end{aligned}
\tag{6.22}
$$

where a sum over fermion species r is understood and I^3_W is the corresponding weak isospin quantum number.

Note that this Lagrangian is invariant under the infinitesimal gauge transformation

$$
\begin{aligned}
A_\mu &\to A_\mu + \partial_\mu \theta^A + ie(W_\mu^+ \theta^- - W_\mu^- \theta^+) &(6.23)\\
Z_\mu &\to Z_\mu + \partial_\mu \theta^Z - ie\frac{\cos\theta_W}{\sin\theta_W}(W_\mu^+ \theta^- - W_\mu^- \theta^+)\\
W_\mu^\pm &\to W_\mu^\pm + \partial_\mu \theta^\pm \mp i\frac{e}{\sin\theta_W}(W_\mu^\pm(\sin\theta_W \theta^A - \cos\theta_W \theta^Z) - (\sin\theta_W u - \cos\theta_W Z_\mu)\theta^\pm)\\
\eta &\to \eta + \frac{e}{2\sin\theta_W\cos\theta_W}\chi\theta^Z + i\frac{e}{2\sin\theta_W}(\phi^+\theta^- - \phi^-\theta^+)\\
\chi &\to \chi - \frac{e}{2\sin\theta_W\cos\theta_W}(v+\eta)\theta^Z + \frac{e}{2\sin\theta_W}(\phi^+\theta^- + \phi^-\theta^+)\\
\phi^\pm &\to \phi^\pm \mp ie\phi^\pm\left(\theta^A + \frac{\sin^2\theta_W - \cos^2\theta_W}{2\cos\theta_W\sin\theta_W}\theta^Z\right) \pm \frac{ie}{2\sin\theta_W}(f+\eta\pm i\chi)\theta^\pm\\
f_i^{f\pm L} &\to f_i^{f\pm L} - ie\left(Q_i^\pm\theta^A + \frac{\sin\theta_W}{\cos\theta_W}\left(Q_i^f \mp \frac{1}{2\sin^2\theta_W}\right)\theta^Z\right)f_i^{\pm L} + \frac{ie}{\sqrt{2}\sin\theta_W}\theta^\pm V_{ij}^\pm f_j^{\pm L}\\
f_i^{fR} &\to -ieQ_i^f\left(\theta^A + \frac{\sin\theta_W}{\cos\theta_W}\theta^Z\right)f_i^{fR}. &(6.24)
\end{aligned}
$$

The $\pm$ index for the left-handed fermion fields counts the isospin directions. The infinitesimal gauge functions θ^α are determined from the underlying weak isospin θ^i and hypercharge θ^Y gauge transformations by

$$
\begin{aligned}
\theta^\pm &= \frac{1}{g'}(\theta^1 \mp i\theta^2)\\
\theta^A &= \frac{1}{g}\cos\theta_W\theta^Y - \frac{1}{g'}\sin\theta_W\theta^3\\
\theta^Z &= \frac{1}{g'}\cos\theta_W\theta^3 + \frac{1}{g}\sin\theta_W\theta^Y.
\end{aligned}
$$

It is now straightforward to upgrade the Lagrangian of the electroweak sector of the standard model (6.22) to the full Lagrangian of the standard model by adding the one for the strong interactions

$$
\begin{aligned}
\mathcal{L}_s &= -\frac{1}{4}G_{\mu\nu}^a G_a^{\mu\nu} - g''\bar{f}_i^{rs}\gamma^\mu\omega^a f_i^{rs}G_\mu^a\\
G_{\mu\nu}^a &= \partial_\mu G_\nu^a - \partial_\nu G_\mu^a + g'' h^{abc}G_\mu^b G_\nu^c,
\end{aligned}
$$

where the generators ω^a and the structure constants h^{abc} belong to the gauge group of the strong interactions, the so-called color group SU(3)[11], and g'' is the corresponding coupling constants. G_μ^a are the gauge fields of the gluons, and the fermions now have also an (implicit) vector structure in the strong-space, making them three-dimensional color vectors or singlets, for quarks and leptons, respectively, as discussed in chapter 5.

[11]Actually SU(3)/Z_3, similar to the case of the weak isospin group.

Note that the number of Higgs fields is larger than the minimal number required: Only two (real) degrees of Higgs are necessary to write down a consistent theory of the weak interactions. However, in this case only one of the weak gauge bosons would become massive. Hence, experiments requires four degrees of freedom. This introduces an additional global symmetry between the four Higgs degrees of freedom, similar to the flavor symmetries of the fermions. It is called a custodial symmetry, as in absence of other interactions it would lead to mass-degenerate W and Z bosons. However, QED breaks this symmetry explicitly, and leads, as has been shown above, to the slight mass difference between W and Z bosons.2

6.11 Anomalies

As noted in section 4.9, it is possible that a symmetry is broken by the quantization. Such a breaking is called an anomalous breaking, mainly for the reason that this was not expected to occur in the beginning. Today, it is clear that such a breaking is deeply linked into the quantization process, and is in most cases also related to the necessity for renormalization.

It is now possible to consider an anomaly for a global or a local symmetry. Anomalies for global symmetries do not pose any inherent problems. Indeed, the axial part $\mathrm{U}_A(1)$ of the chiral symmetry described in section 4.12 is broken anomalously in the standard model. This breaking has directly observable consequences. In that particular case, it leads to a much larger partial decay width of the π^0 into two photons, as it would be without this anomaly. Also the large mass of the η' meson, as indicated in section 5.11 can be shown to originate from the same anomaly.

The situation is much more severe for an anomalous breaking of a local symmetry. If this occurs, this implies that while the classical theory is gauge-invariant, the quantum version would not be so. Hence, the quantum version would depend on the gauge chosen classically, and observables would become gauge-dependent. This would not yield a consistent theory, and would therefore be discarded as a theory of nature. Thus, only theories without local anomalies are considered for particle physics.

Considering theories of the type the standard model, i. e. a number of gauge fields with fermions and scalars, it turns out that the possibility of gauge anomalies is tied to the presence of fermions. Theories with only gauge fields and scalars do not develop anomalies. Furthermore, it can be shown that only theories with parity violations can be affected by local anomalies. Finally, anomalies can only occur for certain gauge groups[12], and not for

[12]In total, the Lie groups $\mathrm{U}(1)$, $\mathrm{SU}(N>2)$, $\mathrm{Sp}(N)$, and $\mathrm{O}(2<N<6)$ are affected.

arbitrary ones, and only for particular assignments of charges to the fermions. However, this affected combination of charge assignments and gauge groups includes the one of the standard model. Thus, in principle, the standard model would be anomalous.

Fortunately, there is an escape route left. Though indeed such anomalies occur, it is possible that if there are more than one anomaly, the consequences of these anomalies cancel each other. If such an anomaly cancellation occurs, the theory becomes anomaly-free, and is again well-defined. Given the gauge groups and charge assignments of the standard model, the condition that all anomalies are canceled and the standard model is consistent if

$$\sum_f Q_f = N_g \left((0-1) + N_c \left(\frac{2}{3} - \frac{1}{3} \right) \right) = N_g \left(-1 + \frac{N_c}{3} \right) = 0,$$

where the sum is over all fermion flavors, quarks and leptons alike, N_g is the number of generations, Q_f is the electric charge of each flavor f, and N_c is the number of colors. As can be seen, the anomalies indeed cancel, and they do so for each generation individually. However, the cancellation requires a particular ratio of the quark electric charges and the lepton electric charges as well as the number of colors. Since the anomalies originate from the parity-violating interaction, all three forces are involved in guaranteeing the anomaly-freedom of the standard model.

This fact has led to many speculations. The most favored one is that this indicates a underlying structure, which provides a common origin to all three force of the standard model. Such theories are known as grand-unifying theories (GUT). Such theories have one underlying gauge group, which would be spontaneously broken at high energies, and therefore appear like three separate forces at low energies. This reasoning is supported by the fact that the running gauge couplings of all three forces all approach a similar size at about 10^{15} GeV, the GUT scale, which suggests a common origin at this energy scale. This will be discussed in detail in section 9.3.

A second explanation is that only a theory like the standard model, with such a fine balance, has all the features necessary to permit sentient life to arise to recognize this feature. In a universe with different laws of nature, where this kind of balance is not met nobody would be there to observe it. This is called the anthropomorphic principle. Though it cannot be dispelled easily, it is not as compelling an argument, as it does not explain why such a universe should exist at all. This problem is often circumvented by the requirement that actually many universes with all kinds of laws of nature exist, and we just happen to be in one where we could exist.

The third possibility is, of course, that all of this is just coincidence, and nature is just this way.

At any rate, future experiments will decided, which of these three options is realized. In fact, experiments have already ruled out the simplest candidates for grand-unified theories, though many remain.

6.12 Landau-poles, triviality, and naturalness

As has been noted, the three gauge couplings of the standard model depend on energy. The strong and weak coupling diminish at high energies, and both interactions are asymptotically free. However, both have very different energy behaviors: The strong interaction becomes stronger towards the infrared, while the weak interaction reaches a maximum, and then diminishes also towards small energies, mostly due to the screening by the Higgs effect. The electromagnetic coupling, however, rises towards large energies.

Assume for a moment that the three couplings unify at some point, as discussed in the previous section 6.11, and that the resulting GUT is asymptotically free. Hence, the high-energy behavior of the gauge interactions becomes well-defined. Then there are still 13 non-gauge interactions in the standard model left: The four-Higgs self-interaction, and the 12 Yukawa couplings between the Higgs and the fermions[13]. As all these interactions are not of a gauge type, they are all not asymptotically free. Hence, all of them rise with energy, though even the largest one, the top-Higgs-Yukawa coupling, very slowly. As a consequence, perturbation theory fails at some scale due to the emergence of Landau poles in these couplings[14].

This rise is strongly dependent on the mass of the Higgs. If the Higgs mass increases, the couplings start to rise quicker. For a Higgs mass of around 1 TeV, perturbation theory will already fail at 1 TeV, as then the four-Higgs coupling develops the Landau pole at the same energy. At the same time a Higgs of this mass will have a decay width of the same order as its mass. This shows that the physics in the Higgs sector is very sensitive to the values of the parameters, in fact quadratically. This is in contrast to the remainder of the standard model. Both the fermion and gauge properties are only logarithmically sensitive to the parameters.

To achieve that the Higgs mass is of the same order as the W and Z mass requires that the parameters in the standard model are very finely tuned, to some ten orders

[13]It is assumed for simplicity here that all neutrinos are massive, though the experimental data are still compatible with one of them being massless.

[14]Occasionally one finds in the literature the argument of perturbativity, i. e. that any decent theory describing high energies must have weak coupling, and be perturbatively treatable. However, there are no reasons for this assumption except technical ones. Hence, such statements express a hope rather than any reasonable physics.

of magnitude. This is the so-called fine-tuning problem. In fact, if random values for the couplings would be taken, the Higgs mass would usually end up at masses orders of magnitudes larger than the W and Z mass. Why this is not the case is called the hierarchy problem: Why is there no large hierarchy between the electroweak scale and the Higgs mass scale? This is sometimes rephrased as the naturalness problem: Without any prior information, it would naively be expected that all parameters in a theory are of the same order of magnitude. This is not the case in the standard model. Why do the values deviate so strongly from their 'natural' value?

There is another problem connected with these questions. If only the Higgs sector is regarded, the theory is called a ϕ^4 theory, as it has the same structure as the Lagrangian (5.5), except that there are four scalar fields. It is found that in such a theory quantum effects always drive the self-interaction term in the potential to zero, i. e. the quantum theory is non-interacting, even if the classical theory is interacting. Theories in which this occurs are called trivial. Such a triviality can be removed, if the theory is equipped with an intrinsic energy cutoff, which introduces an additional parameter, but restricts the validity of the theory to below a certain energy level. It is then just a low-energy-effective theory of some underlying theory.

It is not clear whether this triviality problem persists when the Higgs sector is coupled to the rest of the standard model. If it does, the necessary cut-off for the standard model will again depend sensitively on the mass of the Higgs boson. Practically, this is not a serious issue, as with the present mass of the Higgs of roughly 125 GeV this triviality cut-off can be easily as large as 10^{19} GeV, far beyond the current reach of experiments and observations. Still, it remains as a doubt on how fundamental the standard model, and especially the Higgs particle, really is, in accordance with the problems introduced by the necessity of renormalization as described in section 3.7.

6.13 Baryon-number violation

Our current understanding of the origin of the universe indicates that it emerged from a big bang, a space-time singularity where everything started as pure energy. Why is then not an equal amount of matter and anti-matter present today, but there is a preference for matter? CP violation explains that there is indeed a preference for matter over anti-matter, but the apparent conservation of lepton and baryon number seems to indicate that this is only true for mesons and other states which do not carry either of these quantum numbers. This impression is wrong, as, in fact, there is a process violating baryon (and lepton) number conservation in the standard model. Unfortunately, both this process and CP violation turn out to be quantitatively too weak to explain with our current understanding of the early evolution of the universe the quantitative level of the asymmetry between matter and anti-matter. Thus, the current belief is that so far undiscovered physics is responsible for the missing (large$\sim 10^9$) amount.

It is nonetheless instructive to understand how baryon number violation comes about in the standard model. Lepton number violation proceeds in the same way, but is even more suppressed, due to the much smaller masses.

The basic ingredient is a classical field configuration, this time of Yang-Mills theory. Define the field strength tensors $F_{\mu\nu} = \tau^a F^a_{\mu\nu}$, with τ^a the Pauli matrices - here the weak interactions with the gauge group SU(2) is the interesting one. To proceed further, it is useful to make the formal replacement $it \to t$, which can be undone at the end. This is an analytic continuation from Minkowski space-time to Euclidean space-time, as now all components of the metric have the same sign.

The Jacobi identity (5.15) together with the non-Abelian version of the homogeneous Maxwell equation of QED,

$$\partial_\rho F_{\mu\nu} + \partial_\mu F_{\nu\rho} + \partial_\nu F_{\rho\mu} = 0,$$

yield the Bianchi identity

$$\begin{aligned} D_\rho F_{\mu\nu} + D_\mu F_{\nu\rho} + D_\nu F_{\rho\mu} &= 0, \\ D_\mu F_{\nu\rho} &= D^{ij}_\mu \tau^a_{jk} F^a_{\nu\rho}. \end{aligned}$$

Together, they imply

$$\begin{aligned} 0 &= \frac{1}{2}\epsilon_\sigma{}^{\rho\mu\nu} D_\rho F_{\mu\nu} = D_\rho \tilde{F}_{\sigma\rho} \\ \tilde{F}_{\sigma\rho} &= \frac{1}{2}\epsilon_{\sigma\rho}{}^{\mu\nu} F_{\mu\nu}, \end{aligned}$$

where $\tilde{F}$ is called the dual field-strength tensor. The correctness of these statements follow from the anti-symmetry of the field-strength tensor and the Jacobi identity (5.15). However, the (inhomogeneous) Maxwell equation

$$D_\mu F^{\mu\nu} = 0, \tag{6.25}$$

are therefore trivially solved by (anti-)self-dual solutions

$$F_{\mu\nu} = \pm \tilde{F}_{\mu\nu}, \tag{6.26}$$

as these convert the equation (6.25) into the trivial Bianchi identity. The self-duality equations (6.26) have the advantage of being only first-order differential equations instead of second-order differential equations, and are therefore easier to solve.

Furthermore, classical solutions have to have a finite amount of energy, and therefore their behavior at large distances is constrained. Especially, since the Lagrangian can be written as the sum of the squares of the electric and magnetic field strength, both these fields must vanish. This can only occur if the potential becomes at large distances gauge-equivalent to the vacuum, i. e. it has the form

$$A_\mu^a \tau^a = A_\mu = ig(x)\partial_\mu g^{-1}(x),$$

where $g = g^a \tau_a$ is an arbitrary function. Since this is the second part of the gauge transformation (5.19) in matrix notation, this is just a version of the vacuum $A_\mu = 0$. Since all choices are gauge-equivalent, any choice will do. One possibility which turns out to be technically convenient is

$$\begin{aligned} g(x) &= \frac{x^\mu \tau_\mu}{|x|} \\ \tau_\mu &= (1, i\tau^a). \end{aligned} \qquad (6.27)$$

The simplest extension of this is a multiplication with a function $f(x^2)$ which becomes 1 at large distances,

$$\begin{aligned} A_\mu^a \tau^a &= if(x^2)g(x)\partial_\mu g^{-1}(x) = 2f(x^2)\tau_{\mu\nu}\frac{x^\nu}{x^2} \\ \tau_{\mu\nu} &= \frac{1}{4i}(\tau_\mu \bar{\tau}_\nu - \tau_\nu \bar{\tau}_\mu) \\ \bar{\tau}_\mu &= (1, -i\tau^a). \end{aligned}$$

The matrices $\tau_{\mu\nu}$ are called 't Hooft symbols. Thus, this ansatz mixes non-trivially the weak isospin and space-time.

Plugging this in into the self-duality equation (6.26) yields a first-order differential equation for $f(x^2)$,

$$x^2 \frac{df}{dx} - f(1-f) = 0.$$

The solution to this equation, which can be obtained by separation of variables, is

$$f(x^2) = \frac{x^2}{x^2 + \lambda^2},$$

where λ is an integration constant. The function indeed goes to one at large distances, as required. The structure described by this field configuration is now localized in space-time at the origin, and extended over a range of size λ. Such a localized event in space and time is called an instanton. Solving the equation with the other sign in the self-duality equation (6.26) yields a similar result, though with some small differences, and is called an anti-instanton.

Going back to Minkowski space-time, the field configuration will have a singularity at $x^2 = -\lambda^2$. This is called a sphaleron, and hence a violent event in space-time. Note that the gauge-field strength does not explicitly appear in the calculation. This result can therefore not be obtained perturbatively, and the presence of instantons is a non-perturbative effect.

Seeing now that these indeed create baryon number violation is unfortunately technically very complicated, so here only the most important steps will be sketched. One highly non-trivial, but very fundamental, insight needed is that instantons turn out to be very much connected with the anomalies of section 6.11. In fact, any instanton induces via the global anomaly interactions between fermions. Especially, it can connect fermions of different types, as long as all are charged and all are affected by the anomaly. Especially, this yields that instantons effectively create an interaction which involves, besides the gauge fields also three quarks and one lepton, and thus permits baryon number violation. However, this integration still conserves fermion number, and as a consequence the change in the baryon number must be offset by the same change in lepton number. Still, this implies that reducing the baryon number by one can be offset by a change in lepton number by one, which implies that a proton can be converted, with the involvement of gauge fields, into a positron. Hence, baryon (and lepton number) are not conserved in the standard model.

The effect is, however, small, and suppressed exponentially by $\exp(-c/\alpha_W)$, where c is a number, and α_W is the weak isospin fine-structure constant. Since the latter is small, the suppression is huge, and the life-time of the proton in the standard model exceeds the current upper experimental limit for the decay in any channel of 2.1×10^{28} years by many orders of magnitude. Hence, the baryon number violation in the standard model is not able to explain the fact that so much more baryons than anti-baryons exist. However, this is a statement about the current state of the universe, and especially its temperature, and it may change in earlier times. Especially, it can be shown that the effect becomes exponentially enhanced with temperature. Still, the temperatures necessary to make this a sufficiently effective process have been available for too short a time in the early universe.

6.14 The early universe

The weak interactions, and particularly the Brout-Englert-Higgs effect, play also another important role in the early universe. Just as the magnetization in a magnet can be removed by heating it up, so can the Higgs condensate melt at high temperatures, making the symmetry manifest once more. This process is different in nature from the effectively manifest symmetry at large energies, which is a quantitative effect as all masses become negligible. However, there is not necessarily a phase transition associated with the melting of the condensate. In fact, it is in general not a symmetry restoration, as in case of a global symmetry to be discussed here as well. This is due to the fact that the symmetry is just hidden, not broken. As such, there is no local gauge-invariant order parameter associated with it. Only gauge-dependent order parameters can be local, but in their case the temperature where the symmetry becomes manifest once more is in general gauge-dependent. Only a transition which would be indicated by a non-local gauge-invariant order parameter could in principle mark a true phase transition.

Again, it is quite useful to first study the case of a global symmetry, then of an Abelian local symmetry before going to the electroweak theory.

6.14.1 Global symmetry

A useful starting point is given by the Lagrangian (5.8)

$$\begin{aligned}\mathcal{L} &= \frac{1}{2}\partial_\mu\eta\partial^\mu\eta-\frac{1}{2}(6\lambda f^2-\mu^2)\eta^2+\frac{1}{2}\partial_\mu\chi\partial^\mu\chi-\frac{1}{2}(2\lambda f^2-\mu^2)\chi^2\\ &\quad -\sqrt{2}\lambda f\eta(\eta^2+\chi^2)-\frac{1}{4}\lambda(\eta^2+\chi^2)^2-\mu^2f^2+\lambda f^4. \end{aligned} \tag{6.28}$$

In this case the explicit zero-energy contribution is kept for reasons that will become apparent shortly, but will be essentially the same as when treating non-relativistic Bose-Einstein condensation. Only terms linear in the fields have been dropped, as they will not contribute in the following. The situation is similar as before, but now the condensate f has not been specified by the minimization of the classical potential, but is kept as a free quantity, which will take its value dynamically.

To investigate the thermodynamic behavior it is useful to analyze the thermodynamic potential Ω in analogy to the non-relativistic case as

$$\Omega(T,f)=-P(T,f)=-T\ln\frac{Z}{V},$$

where P is the pressure, T the temperature, and V the volume. Z is the partition sum. For the following purposes, it is sufficient to use the so-called mean-field approximation. In this case, the interaction terms are first expanded around f, and higher-order terms which involve more than two fields are neglected. In the Lagrangian (6.28), the first step has already been done.

Without going into the technical details, the thermodynamic potential can be evaluated directly. It reads

$$\begin{aligned}\Omega(T,f) &= -\mu^2 f^2+\lambda f^4 \\ &\quad+\int\frac{d^3p}{(2\pi)^3}\left(\frac{\omega_1^2+\omega_2^2}{2}+T\left(\ln\left(1-e^{-\frac{\omega_1}{T}}\right)+\ln\left(1-e^{-\frac{\omega_2}{T}}\right)\right)\right) \\ \omega_1 &= \sqrt{6\lambda f^2-\mu^2+p^2}=\sqrt{m_\eta^2+p^2} \\ \omega_2 &= \sqrt{2\lambda f^2-\mu^2+p^2}=\sqrt{m_\chi^2+p^2}.\end{aligned} \tag{6.29}$$

The frequencies ω consists of the momenta and the masses of the particles after hiding the symmetry, which is dependent on the value of the condensate f. There are three contributions. The first outside the integral is the classical contribution. The second are the first two terms inside the integral. They are the contributions from quantum fluctuations. The third term represents thermal fluctuations.

To recover the results from section 5.7, the second term must be neglected and the zero-temperature limit taken. This yields

$$\Omega(0,f)=-\mu^2f^2+\lambda f^4.$$

As in section 5.7, this potential has a minimum at non-zero f, $f=\mu^2/(2\lambda)$. Inserting this into the Lagrangian (6.28) makes it equivalent to (5.8), with a massive Higgs boson and a massless Goldstone boson.

Something new happens at finite temperature. At small temperature, little other happens than that it is possible to excite Higgs bosons or Goldstone bosons, which then form a thermal bath of non-interacting bosons, and the total pressure is just the sum of their respective pressures. However, the value of f will become temperature-dependent: At each temperature it will take the value which minimizes the thermodynamic potential.

When going to higher temperatures, it is useful to make a high-temperature expansion for the thermodynamic potential. High temperature requires here T to be larger than the scale of the zero-temperature case, which is given by the condensate, which is of order $\mu/\sqrt{\lambda}$. In this case, it is possible to obtain an expansion for Ω. The leading terms up to order $\mathcal{O}(1)$ are given by

$$\Omega(T,f)=\lambda f^4+\left(\frac{1}{3}\lambda T^2-\mu^2\right)f^2-\frac{\pi^2}{45}T^4-\frac{\mu^2T^2}{12}. \tag{6.30}$$

Note that at very large temperatures only the term $\pi^2T^4/45$ is relevant, which is precisely the one of a free non-interacting gas of two boson species, a Stefan-Boltzmann-like behavior.

This results exhibits one interesting feature. The term of order f^2 has a temperature-dependent coefficient, which changes sign at[15] $T_c^2 = 3\mu^2/\lambda$. As a consequence, the shape of the thermodynamic potential as a function of f changes. Below T_c, it has a minimum away from zero, as at zero temperature. With increasing temperature, this minimum moves to smaller and smaller values, and arrives at zero at T_c. Hence, at T_c, the value of f changes from a non-zero to a zero value, and the symmetry becomes manifest once more. Above T_c, the minimum stays at zero, and for all higher temperatures the symmetry is manifest.

Replacing f with its temperature-dependent value in (6.30) yields the expressions

$$\begin{aligned}
\Omega_{T<T_c} &= \tfrac{\mu^2T^2}{12} - \left(\tfrac{\pi^2}{45}+\tfrac{\lambda}{36}\right)T^4 \overset{T=T_c}{=} -\frac{\pi^2\mu^2}{5\lambda^2} \\
\Omega_{T>T_c} &= \tfrac{\mu^4}{4\lambda} - \tfrac{\pi^2}{45}T^4 - \tfrac{\mu^2T^2}{12} \overset{T=T_c}{=} -\frac{\pi^2\mu^2}{5\lambda^2},
\end{aligned}$$

which coincide at T_c. Also their first derivatives with respect to the temperature are equal at T_c

$$\begin{aligned}
\frac{d\Omega_{T<T_c}}{dT} &= -(8\pi^2T^2+10\lambda T^2-15\mu^2)\tfrac{T}{90} \overset{T=T_c}{=} -\frac{8\pi^2+5\lambda}{\sqrt{300}}\sqrt{\frac{\mu^2}{\lambda}} \\
\frac{d\Omega_{T>T_c}}{dT} &= -(8\pi^2T^2+15\mu^2)\tfrac{T}{90} \overset{T=T_c}{=} -\frac{8\pi^2+5\lambda}{\sqrt{300}}\sqrt{\frac{\mu^2}{\lambda}},
\end{aligned}$$

but their second derivatives are not

$$\begin{aligned}
\frac{d^2\Omega_{T<T_c}}{dT^2} &= \tfrac{\mu^2}{6} - (4\pi^2+5\lambda)\tfrac{T^2}{15} \overset{T=T_c}{=} -\frac{(25\lambda+24\pi^2)\,\mu^2}{30\lambda} \\
\frac{d^2\Omega_{T>T_c}}{dT^2} &= -\tfrac{8\pi^2T^2+5\mu^2}{30} \overset{T=T_c}{=} -\frac{(5\lambda+24\pi^2)\,\mu^2}{30\lambda}.
\end{aligned}$$

Thus, a phase transition of second order occurs at T_c.

As stressed previously repeatedly, it is possible that quantum effects could modify the pattern considerably or even melt the condensate. It is therefore instructive to investigate the leading quantum corrections to the previous discussion.

This is also necessary for another reason. If the symmetry becomes manifest once more at large temperatures, the mass of Higgs-like excitations become tachyonic, indicating a flaw of the theory. That can be seen directly by reading off the condensate-dependent masses of the excitations,

$$\begin{aligned}
m_\eta^2 &= 6\lambda f^2 - \mu^2 = -\mu^2\theta(T-T_c) + (2\mu^2-\lambda T^2)\theta(T_c-T) \\
m_\chi^2 &= 2\lambda f^2 - \mu^2 = -\mu^2\theta(T-T_c) - \frac{\lambda T^2}{3}\theta(T_c-T).
\end{aligned}$$

[15]Note that strictly speaking using the high-temperature expansion at this temperature is doubtful. For the purpose here it will be kept since it makes the mechanisms more evident then the rather technical calculations necessary beyond the high-temperature expansion. The qualitative outcome, however, is not altered, at least within the first few orders of perturbation theory.

Furthermore, also the Goldstone theorem is violated, as the mass of the Goldstone boson χ is no longer zero[16]. Both problems are fixed by quantum corrections, demonstrating the importance of quantum fluctuations even in the high-temperature phase.

In the expression for the free energy (6.29) the zero-point energy, and thus the quantum fluctuations have been neglected. Including them yields the result

$$\begin{aligned}\Omega(T,f) &= -\frac{\pi^2}{45}T^4-\frac{\mu^2T^2}{12}-\frac{(m_\eta^3+m_\chi^3)T}{12}+\frac{\mu^4}{32\pi^2}\ln\frac{8\pi^2T^2e^{-2\gamma+\frac{3}{2}}}{\mu^2}\\ &\quad -\mu^2f^2\left(1+\frac{\delta\xi}{\lambda}+\frac{\lambda}{4\pi^2}\ln\frac{8\pi^2T^2e^{-2\gamma+1}}{\mu^2}-\frac{\lambda T^2}{3\mu^2}\right)\\ &\quad +\lambda f^4\left(1+\frac{\delta\xi}{\lambda}+\frac{5\lambda}{8\pi^2}\ln\frac{8\pi^2T^2e^{-2\gamma+1}}{\mu^2}\right).\end{aligned}$$

Herein γ is the Euler number. The term $\delta\xi$ is a quantum correction, which can be fixed, e. g. by measuring T_c or by other experimental input, as can be seen from the new expression for T_c where f vanishes,

$$T_c^2=\frac{3\mu^2}{\lambda}\left(1+\frac{\delta\xi}{\lambda}+\frac{\lambda}{4\pi^2}\ln\frac{24\pi^2e^{-2\gamma+1}}{\lambda}\right).$$

Here, only the qualitative result is interesting, and its precise numerical value of no importance.

To obtain the corrections for the masses, it is necessary to calculate the corresponding quantum corrections. Without going into the details, the complete masses to this order are

$$\begin{aligned}m_\eta^2 &= 2\mu^2\left(1-\frac{\lambda T^2}{3\mu^2}\right)\theta(T_c-T)+\frac{1}{3}\lambda\left(T^2-\frac{3\mu^2}{\lambda}\right)\theta(T-T_c)\\ m_\chi^2 &= \frac{\lambda}{3}\left(T^2-\frac{3\mu^2}{\lambda}\right)\theta(T-T_c).\end{aligned}$$

These results yield a number of interesting observation. First, since T_c is larger[17] than $3\mu^2/\lambda$, the mass of the Higgs is always positive, stabilizing the system. Secondly, in this case the mass of the Goldstone boson is always zero below the phase transition temperature, in agreement with the Goldstone theorem. Above the phase transition, the masses of both particle degenerate, and the symmetry is manifest once more also in the spectrum.

[16] In a full quantum treatment, the role of the Goldstone boson could be played at finite temperature by some composite excitation instead. However, at the current, so-called, mean-field level no such excitations are available, and thus the Goldstone theorem is violated.

[17] It is not obvious that $\delta\xi$ cannot be negative and large, thus making the improved estimate for T_c smaller than before. However, it turns out not to be the case at this order for any renormalization prescription.

These properties are generic for symmetries hiding by a condensate which thaws with increasing temperature. Also that the mean-field approximation is in general insufficient is a lesson which should be kept duly in mind. Of course, at the present time much more sophisticated methods are available to treat this problem, though they are in general very complicated.

6.14.2 Abelian case

As was visible in the previous case of the global symmetry, quantum fluctuations are important to obtain a consistent result. Again going through the mean-field calculations (introducing only a mean-field for the Higgs fields η condensate, but neither for the pseudo-Goldstone boson χ nor for the photon), it is possible to obtain once more a high-temperature expansion for the free energy. The full expression at mean-field level then reads

$$\Omega(f,T)=\lambda f^4+\left(\left(\frac{\lambda}{3}+\frac{e^2}{4}\right)T^2-\mu^2\right)f^2-\frac{\mu^2T^2}{12}-\frac{2\pi^2}{45}T^4,$$

where e is now the electromagnetic coupling constant. First, the Stefan-Boltzmann contribution proportional to T^4 is now twice as large, as the photon also contributes two scalar degrees of freedom. Furthermore, the leading high-temperature behavior is not altered otherwise by the presence of the interaction. At the level of the mean-field approximation there is still only a non-interacting gas of two scalars and two photon polarizations left. However, the term crucial for the phase transition, the one of order T^2, is affected by the interactions. Thus, the interactions have an influence on the phase transition. At mean-field level they just shift the phase transition temperature.

Beyond mean-field, their impact is more relevant. At one-loop level, the phase structure becomes dependent on the relative size of λ and e.

As long as λ is larger[18] than e, the situation is found to be in agreement as when the photon field would be absent. In particular, the condensate melts at a (now also e-dependent) critical temperature, and above the phase transition both the scalars and the photon become massless again.

If λ is smaller than e, the critical temperature becomes lower than the effective mass of the photon. As a consequence, the second-order phase transition may in fact become first order. Even more drastic, if λ is below $3e^4/(32\pi^2)$, the phase transition temperature decreases to zero, and spontaneous breaking is not occurring at all. In that case, the stronger photon-scalar interactions stabilize the vacuum, and no condensate can form.

[18] If e and λ are approximately equal, even higher order corrections can have a qualitative impact.

An important insight is found in the case of $\mu = 0$. At first sight no condensation is possible. However, the interactions mediated by the photon may still be sufficiently strong and attractive enough that the scalars condense, yielding the same physics as if μ would not be zero, and λ would be sufficiently large. In this case, the condensation is a genuine quantum effect, as only due to (one-)loop quantum corrections spontaneous condensation of the Higgs occurs.

6.14.3 The electroweak case

Also in the electroweak case it is useful to use the Lagrangian in the 't Hooft gauge, given by (6.22). Though straight-forward, the calculation in the present case is much more cumbersome than in the previous cases. For the present purpose, the masses of the fermions (and thus the CKM matrices) can be neglected. The imprecision of this is comparable to the effects of going to the next order. The final result at mean-field level is

$$\Omega(f,T) = -\frac{427}{360}\pi^2 T^4 - \frac{f^2}{2}\left(\mu^2 - \frac{T^2}{4}\left(2\lambda + \frac{3g^2}{4} + \frac{g'^2}{4}\right)\right) + \frac{\lambda f^4}{4} + \frac{\mu^4}{4\lambda} - \frac{\mu^2 T^2}{6}.$$

Extremalizing this expression with respect to f yields the critical temperature for the electroweak standard model in this approximation as

$$T_c^2 = \frac{4\mu^2}{2\lambda + \frac{3g^2}{4} + \frac{g'^2}{4}}.$$

The temperature dependence of the condensate f and the pressure then read, respectively,

$$\begin{aligned} f^2 &= \frac{\mu^2}{\lambda}\left(1 - \frac{T^2}{T_c^2}\right)\theta(T_c - T) \\ P &= \begin{cases} \frac{427\pi^2 T^4}{360} + \frac{\mu^2}{4\lambda}\left(1 - \frac{T^2}{T_c^2}\right)^2 + \frac{\mu^2 f^2 T^2}{6} - \frac{\mu^4}{4\lambda}, & T \leq T_c \\ \frac{427\pi^2 T^4}{360} + \frac{\mu^2 T^2}{6} - \frac{\mu^4}{4\lambda}, & T \geq T_c \end{cases}. \end{aligned}$$

The corresponding phase transition is at mean-field level thus of second order, as the second derivative of the pressure exhibits a discontinuity. For a Higgs mass of 100 GeV the critical temperature is about 200 GeV. For the actual value of the Higgs mass of 126 GeV, it is only slightly higher. Thus the transition temperature is of the same order as the Higgs condensate, and about three orders of magnitude larger than the corresponding temperature in QCD.

Note that all problems with consistency of the mean-field approach pertain also to the full electroweak standard model. Therefore, for a consistent treatment at least leading order corrections have to be included. As previously, they do not change the phase transition temperature, but may change its order. In fact, in the standard model the transition becomes a crossover.

The relevance of such a temperature is only given in the early universe, or perhaps during a supernova collapse to a black hole. Here, the more certain case of the early universe will be treated.

To assess the relevance for the early universe it is necessary to add equations which describe its development. For the present purpose simplified versions of the Einstein equations are sufficient. A more detailed discussion of this will be given in section 8.4.1. Adding energy conservation gives

$$\left(\frac{dR}{dt}\right)^2 = \frac{8\pi G}{3}\epsilon R^2 \tag{6.31}$$

$$\frac{d(\epsilon R^3)}{dR} = -3PR^2, \tag{6.32}$$

where G is Newton's constant, ϵ is the energy density, R is a quantity characterizing the size of the universe, t the proper time, and P is the pressure. To close the equations, an equation of state is necessary, which is given as a function of the pressure by thermodynamic relations as

$$\epsilon = -P + T\frac{\partial P}{\partial T}.$$

Since the sicknesses of the mean-field approximation are not too problematic for this estimate, it is sufficient to use it for obtaining the corresponding energy density as

$$\epsilon = \begin{cases} \left(\frac{1281\pi^2}{360} + \frac{\mu^4}{4\lambda T_c^4}\right)T^4 + \left(1 - \frac{3\mu^2}{\lambda T_c^2}\right)\frac{\mu^2 T^2}{6}, & T \leq T_c \\ \frac{1281\pi^2}{360}T^4 + \frac{\mu^2 T^2}{6} + \frac{\mu^4}{4\lambda}, & T \geq T_c \end{cases}.$$

Rewriting equation (6.32) in terms of the temperature yields the ordinary differential equation

$$R\frac{d\epsilon}{dT} + 3\frac{dR}{dT} = -3P.$$

Imposing as a boundary condition that R should be one at the phase transition yields

$$R^3 = \begin{cases} \frac{T_c}{T}\frac{T_c^2-b^2}{T^2-b^2}, & T \leq T_c \\ \frac{T_c}{T}\frac{T_c^2+a^2}{T^2+a^2}, & T \geq T_c \end{cases}$$

$$a^2 = \frac{30\mu^2}{427\pi^2}$$

$$b^2 = \frac{a^2 T_c^2(1-r)}{a^2 + rT_c^2}$$

$$r = \frac{4\lambda}{6\lambda + \frac{9g^2}{4} + \frac{3g'^2}{4}}.$$

For a Higgs mass about 100 GeV the characteristic parameters are $r = 0.22$, $a = 6$ GeV and $b = 10$ GeV. Thus, the dominant behavior is that R behaves like T_c^3/T^3, up to some small modifications close to the phase transition, and thus drops essentially in the electroweak domain.

An interesting consequence is obtained if R^3 is multiplied by the entropy

$$s = \frac{\partial P}{\partial T} \sim T^3.$$

An elementary calculations yields thus that sR^3 is constant. Since s is in units inverse length cubed, this is just the statement that entropy is conserved since R only describes the expansion of a unit length over time. Hence, the electroweak interactions at mean-field level conserve entropy. In particular, this is a consequence of the second order phase transition, which is not permitting latent heat or supercooling. Finally, inserting the numbers shows that R increases somewhat slower around the phase transition. Thus, the expansion of the universe slows down during the electroweak phase transition.

Of course, all of this is just an estimate. That it can never be fully correct is seen by the fact that R diverges at the finite temperature $T = b$, much above the QCD phase transition (and nowadays) temperature. This is an artifact of the high-temperature expansion involved. To obtain the correct behavior down to the QCD phase transition would require more detailed calculations. This would also uncover other interesting features addressed later.

Chapter 7

Gravity

Gravity, as encoded in general relativity, is besides the standard model the other basic theory known so far. In contrast to the standard model, it contains actually very few structures. As such, despite its importance, it actually encodes relatively few physical phenomena when considered alone. Only by coupling it to matter governed by other interactions it becomes much more richer.

7.1 Space-time and the metric

The basic object of general relativity is the symmetric metric tensor $g_{\mu\nu}$. It describes the invariant length-element ds by

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu.$$

This encodes the invariant distance between two events. It is as such a genuine physical quantity. Coordinates are, in fact, not, as will be seen later.

The inverse of the metric is given by the contravariant tensor

$$g^{\mu\nu} g_{\nu\lambda} = \delta^\mu_\lambda.$$

As a consequence, for any derivative δ

$$\delta g^{\mu\nu} = -g^{\mu\lambda} g^{\nu\rho} \delta g_{\lambda\rho} \tag{7.1}$$

holds. The metric is assumed to be non-vanishing and has a signature such that its determinant is negative,

$$g = \det g_{\mu\nu} < 0.$$

The covariant volume element dV is therefore given by

$$\begin{aligned} dV &= \omega d^4x \\ \omega &= \sqrt{-g} = \sqrt{-\det g_{\mu\nu}} > 0, \end{aligned}$$

implying that h is real (hermitian), and has derivative

$$\delta\omega = \frac{1}{2}\omega g^{\mu\nu}\delta g_{\mu\nu} = -\frac{1}{2}\omega^g_{\mu\nu}\delta g^{\mu\nu} \tag{7.2}$$

as a consequence of (7.1).

The most important concept of general relativity is the covariance (or invariance) under a general coordinate transformation $x_\mu \to x'_\mu$ (diffeomorphism) having[1]

$$\begin{aligned} dx'^\mu &= \frac{\partial x'^\mu}{\partial x^\nu}dx^\nu = J^\mu_\nu dx^\nu \\ \det(J) &\neq 0, \end{aligned} \tag{7.3}$$

where the condition on the Jacobian J follows directly from the requirement to have an invertible coordinate transformation everywhere. Scalars $\phi(x)$ are invariant under such coordinate transformations, i. e., $\phi(x) \to \phi(x')$. Covariant and contravariant tensors of n-th order transform as

$$T'_{\mu\ldots\nu}(x') = \frac{\partial x_\mu}{\partial x'_\alpha}\ldots\frac{\partial x_\nu}{\partial x'_\beta}T_{\alpha\ldots\beta}(x)$$
$$T'^{\mu\ldots\nu}(x') = \frac{\partial x'^\mu}{\partial x^\alpha}\ldots\frac{\partial x'^\nu}{\partial x^\beta}T^{\alpha\ldots\beta}(x)$$

respectively, and contravariant and covariant indices can be exchanged with a metric factor, as in special relativity. As a consequence, the ordinary derivative ∂_μ of a tensor A_ν of rank one or higher is not a tensor. To obtain a tensor from a differentiation the covariant derivative must be used

$$\begin{aligned} D_\mu A_\nu &= \partial_\mu A_\nu - \Gamma^\lambda_{\mu\nu}A_\lambda \\ \Gamma^\lambda_{\mu\nu} &= \frac{1}{2}g^{\lambda\sigma}(\partial_\mu g_{\nu\sigma} + \partial_\nu g_{\mu\sigma} - \partial_\sigma g_{\mu\nu}), \end{aligned} \tag{7.4}$$

where Γ are the Christoffelsymbols. As a consequence, covariant derivatives no longer commute, and their commutator is given by the Riemann tensor $R_{\lambda\rho\mu\nu}$ as

$$\begin{aligned} \left[D_\mu, D_\nu\right] A^\lambda &= R^\lambda_{\rho\mu\nu}A^\rho \\ R^\lambda_{\rho\mu\nu} &= \partial_\mu\Gamma^\lambda_{\nu\rho} - \partial_\nu\Gamma^\lambda_{\nu\rho} + \Gamma^\lambda_{\mu\sigma}\Gamma^\sigma_{\nu\rho} - \Gamma^\lambda_{\nu\rho}\Gamma^\sigma_{\mu\rho}, \end{aligned}$$

which also determines the Ricci tensor and the curvature scalar

$$\begin{aligned} R_{\mu\nu} &= R^\lambda_{\nu\mu\lambda} \\ R &= R^\mu_\mu, \end{aligned}$$

[1] This is actually only making translations local. Rotations and boosts are a more involved. This is beyond the scope of this lecture.

respectively.

The most subtle consequence of the possibility of local coordinate transformations (7.3) is probably that coordinates are no longer physical in the same sense as in special relativity. Since local coordinate transformations can be done by anyone differently everywhere, they depend now on this choice. Thus, an information like "This object is at $x = 1$ meter" make no longer sense. Only statements like "The invariant distance ds between these two objects is 1 m" does.

This has another, quite unexpected, consequence. Fourier transformations to momentum space involve an expression $g_{\mu\nu}x^\nu p^\mu$. However, because this is not an invariant distance, and x can change by a diffeomorphism transformation, so needs p to change to make this invariant. As a consequence, the 4-momentum depends on the choice of coordinate system. Especially, energy is no longer a unique concept, and can be changed at will[2]. Energy and momentum are thus no longer valid physical concepts. It is a hard problem to find suitable replacements in arbitrary situations.

Fortunately, for special cases, there can be an escape. Most often, gravitational problems can be considered to happen in a finite volume of space-time, and at asymptotic distances space-time becomes flat again. If this is the case, the total energy and momentum inside the volume are again valid physical concepts. This is what is used to define, e. g., the mass of a black hole, as will be seen later.

7.2 Dynamics and the Einstein equations

These definitions and considerations are sufficient to write down the basic dynamical equation of general relativity, the Einstein equation

$$R_{\mu\nu} = \frac{1}{2}g_{\mu\nu}R + g_{\mu\nu}\Lambda = -\kappa T_{\mu\nu}. \tag{7.5}$$

Herein Λ is the cosmological constant, which will play an important role later on. It is, besides the Newton constant, encoded in $\kappa = 16\pi G_N$, a second constant, which needs to be specified. Conceptually, both constants are at the same level, and play the same role as, e. g., the electric charge. They need to be measured, and are inputs to the theory. This equation is essentially the equivalent of Newton's law for the metric: It provides it time-dependence. The right-hand side is the energy-momentum tensor, which contains everythig having to do with matter.

The Einstein equation (7.5) can be derived as the Euler-Lagrange equation from the

[2]At first sight paradoxical situations like the Unruh effect originate from this.

Lagrangian

$$\mathcal{L} = \omega \left(\frac{1}{2\kappa} R - \frac{1}{\kappa} \Lambda + \mathcal{L}_M \right),$$

where $\mathcal{L}_M$ is the matter Lagrangian yielding the covariantly conserved energy momentum tensor $T_{\mu\nu}$

$$T_{\mu\nu} = \left(-\eta_{\mu\nu} \mathcal{L} + 2 \frac{\delta \mathcal{L}_M}{\delta g^{\mu\nu}} (g_{\mu\nu} = \eta_{\mu\nu}) \right), \tag{7.6}$$

where $\eta_{\mu\nu}$ is the Minkowski metric. Thus, non-gravitational interactions and matter enter the theory in the form of this energy-momentum tensor. This is a the same concept as for the current in Maxwell's equations.

While there is a lot of technical details underneath to fully appreciate the structure and consequences of general relativity, this is actually all that is necessary for it. Thus, despite the enormous complexity of phenomena, the theory at a fundamental level is relatively straightforward. What makes it relatively hard to obtain physics from it is the diffeomorphism invariance. It is relatively straightforward to derive results in a fixed coordinate system. What is hard is to derive these results in a diffeomorphism-invariant way, and understand them. The absence of everyday concepts like energy and momentum is often at the root of this. Nonetheless, in the following a couple of important results will be briefly outlined, though not derived in detail.

7.3 Black holes

Probably one of the most impressive predictions of general relativity is the existence of black holes. A black hole is a concentration of mass such that space-time is sufficiently distorted that the world lines of particles, which approach the black hole below a critical distance, the so-called last stable orbit, will be focused on the so-called event horizon. This is true even for massless particles. As a consequence, they can never escape a black hole's gravity well. However, aside from the effect of tidal forces on an extended object, these particles experience a normal movement. Observers at long spatial distances, however, will stop to be able to receive information from such particles once they got close enough. They are therefore lost to the outside world.

The simplest way of presenting a black hole is the so-called Schwarzschild metric

$$ds^2 = -\left(1 - \frac{2a}{r}\right) dt^2 + \frac{dr^2}{1 - \frac{2a}{r}} + r^2 (d\theta^2 + \sin^2\theta d\phi^2). \tag{7.7}$$

If one integrates the energy-momentum tensor given by this metric, i. e. the left-hand side of the Einstein equation, this yields that a can be interpreted as the total mass of the black hole.

It appears at first sight that there is a singularity at $r = 2m$, which identifies the event horizon, and this singularity would be the reason for the strange effects of the black hole. This is, however, not the case. This so-called coordinate singularity is an artifact of the choice of coordinates. It can be eliminated by a different coordinate choice, e. g. the so-called Kruskal coordinates, though these are cumbersome. When considering quantities, which are independent of the choice of the coordinate system, the true nature of the situation becomes better visible. A particular useful example is the curvature scalar, which turns out to be

$$R = 8\pi a\delta(r),$$

which indicates a vanishing curvature except at the center of the black hole. Another option is the so-called Kretschmann scalar

$$K = R^{\alpha\beta\gamma\delta}R_{\alpha\beta\gamma\delta} = \frac{48G_N^2a^2}{r^6}$$

which is also coordinate-independent, and shows a singular behavior only at the origin. Thus, the black hole is actually not much more pathological than a point charge in electrodynamics or a point mass in Newtonian mechanics. The effect of the horizon is merely telling that close to large masses light bending can become so strong that the only possible world-lines are focused on the event horizon. In fact, far away from an object the behavior is always like (7.7). The only question remaining is whether the density distribution is such that the objects extends beyond $r = 2m$, and thus no event horizon develops, or not.

Still, just like the singularities encountered with point objects in classical physics this result indicates that also general relativity breaks down eventually, when moving too close to the center of a black hole. At that point, supposedly, quantum gravity becomes important, just as quantum effects become important for point charges. Whether any of this is relevant already outside the event horizon is yet unclear. Something more will be said about this in section 9.5.1.6.

The solution presented here is the so-called Schwarzschild solution, which is essentially a black hole at rest. This can be upgraded to a spinning black hole, the so-called Kerr black hole. This has very interesting features for the structure of space-time around the event horizon. Moreover, the spin and the mass turn out to be related to some extent. But this leads beyond the scope of this lecture.

As a final remark, it should be added that a black hole in general relativity can at most carry some further global charges, like electric charges, but not any other microstructure. This is known as the no-hair theorem. Furthermore, in four space-time dimensions black hole solutions always show such an ellipsoidal shape. In other dimensions, this may be different. Most infamous is arguably the black Saturn in five dimensions, where a spherical black hole is orbited by a black hole donut, and this is a stable configuration. This would not be possible in four dimensions.

7.4 Compact stellar objects

Given any matter, which only interacts gravitationally, it will always collapse eventually into a black hole. This process can be delayed by having kinetic energy, but by gravitational interactions, this can be dissipated into gravitational waves, as discussed in section 7.5. Of course, such a process can take much longer than the age of the universe.

In addition, non-gravitational interactions can stabilize matter against gravitational collapse, essentially arbitrarily long. To give an example for this, consider a spherical symmetric ideal fluid blob of matter governed by an equation of state $p(\rho)$, i. e. of pressure as a function of the energy density ρ. The mass as a function of radius in equilibirum is then determined by the Tolman-Oppenheimer-Volkoff equation

$$\frac{dp(r)}{dr} = -\frac{(p(\rho(r)) + \rho(r)](M(r) + 4\pi r^3 p(r))}{r\ (r - 2M(\rho(r)))},$$

where $M(r)$ is the mass-energy enclosed within the radius r,

$$M(r) = 4\pi \int\limits_0^r \rho(r')\ r'^2\ dr'.$$

These equations, under the stated assumptions, can be derived from the Einstein equation (7.5).

This coupled set of equations can be reformulated as an initial value problem, for which the density (or pressure) at the center needs to be fixed as an integration constant. The integration is then done until the pressure drops to zero. This gives the mass and radius of the object as a function of the central density. This can be inverted to provide the central density as a function of the measurable mass of the object.

Of course, everything hinges on the accurate knowledge of the equation of state. For ordinary matter, this equation yields essentially objects like planets or stars. It becomes more interesting for matter, in which quantum effects play a role. The predominant effect here is the Pauli principle, as it provides a hard core repulsion. As long as electrons are there, this yields white dwarfs. However, once the gravitational mass becomes large enough to push the electrons into the nuclei, the relevant degrees of freedom become neutrons. This yields neutron stars, objects of about one to a little more than two solar masses, and a radius of order ten kilometers. What actually the maximum mass for a neutron star is, is not known. For this the current understanding of the equation of state of QCD is not yet well enough, as discussed already in section 5.18. It is extremely hard to find a consistent one to push the maximum mass above some 2.5 solar masses. Since so far also observations of neutron stars do not exceed, within errors, such a limit, and are more suggestive of 2.2 solar masses as a maximum, this seems to be consistent.

Whether in a further collapse to even more fundamental structures like quarks, or unknown fermions, is possible, is not yet known. There are speculations about objects like strange stars from strange quark/hypernuclei-dominated matter to such made up only of quarks. But it is not yet clear whether this thermodynamically feasible, and no observations have been made strongly suggesting this. Even more interesting would be options in which unknown new fermions would create even further compact stellar objects. But again, no observational hints in this direction exist.

If the equation of state cannot yield a solution with a surface outside the event horizon, this would yield a black hole. However, before this the assumption of equilibirum breaks down, and more involved equations need to be used, though they ultimately yield the Schwarzschild or Kerr solution.

While these have been the equilibrium solutions, the reality is very often out of equilibrium. However, since on suitable timescales all systems tend towards equilibirum, these solutions are the ultimate fate. While objects like planets or brown dwarfs are already close enough to the equilibirum, stars undergo an evolution, which ends up eventually either as white dwarf, neutron stars or black holes, in order of ascending masses. The transition in the late stage for neutron stars and black hole is violent, known as supernova. In addition, there is a gap between the fate of neutron star and black hole where the supernova just tears the star apart, as it seems, leaving no remnant but gas.

The important insight in these processes is that, while fusion can create up to iron during the burning of stars, nothing beyond this point can happen. Just because iron is the most stable nucleus. However, the temperatures occurring in supernovas can exceed 10-20 MeV, thus allowing thermodynamic formation of even heavier nuclei. This becomes the realm of nuclear astrophysics. In fact, in the formation of neutron stars the interactions become so violent that ultimately even neutrinos become trapped, accelerating the formation of elements. And, in fact, cooling by neutrinos are an important aspect in the later stages of neutron star formation. Thus, neutrino astronomy has become an important part in nova-type events.

It was long believed that (super/hyper)novas produced all these heavy elements. The detailed reaction chains occurring are a separate topic in itself, and involve a substantial amount of so-called nuclear astrophysics. However, in the last decade or so it became clear that ejecta in neutron star mergers are an important source for heavy elements as well, and can contribute about half of it. This insight has been slowly building, and involved elements like the unexpected high rates of double and triple star systems, and the surprising large population of cases where (at least) two stars survived becoming supernovas. Recent first observations of neutron star mergers by gravitational waves with additional optical observations have confirmed this scenario.

7.5 Gravitational wave astronomy

At the mathematical level, general relativity is a very involved theory, much more complicated than electromagnetism. However, these complications are only relevant in the case of strong gravitational fields. In most instances in the universe, these are relatively weak.

In such a case, it is viable to consider general relativity only as a weak perturbation on top of special relativity. This is achieved by splitting the metric as

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}, \tag{7.8}$$

where $\eta_{\mu\nu}$ is the usual Minkowski metric, and $|h_{\mu\nu}| \ll 1$. There is no similar situation in electromagnetism, because there zero field is the baseline, and one can just assume that the field strengths themselves[3] are small for such an approach. Here, this is quite different, and the situation is more akin to the Brout-Englert-Higgs effect in section 6.5.

Even then, the dynamical equation (7.5) remains cumbersome, even if it is linear,

$$\partial^2 h_{\mu\nu} + \partial_\mu \partial_\nu h_\lambda^\lambda - \partial_\lambda \partial_\nu h_\mu^\lambda - \partial_\lambda \partial_\mu h_\nu^\lambda - \eta_{\mu\nu} \partial^2 h_\lambda^\lambda + \eta_{\mu\nu} \partial_\lambda \partial_\sigma h^{\lambda\sigma} = -8\pi G_N T_{\mu\nu}. \tag{7.9}$$

However, for vanishing energy-momentum tensor, it is solved by a plane-wave-like solution,

$$h_{\mu\nu} = \Re \left(\epsilon_{\mu\nu} e^{i\eta_{\mu\nu} k^\mu x^\nu} \right)$$

where the polarization tensor satisfies $\epsilon k = 0$ and it is traceless. Thus, there are at most five independent components. A number of further constraints show that of these actually only two are independent, so the physical polarizations are two, just as for electromagnetic waves. This coincidence stems actually from the dimensionality of space-time, and would not happen for a different number of dimensions. Because the polarization is a tensor, it comes out that at least a quadrupole is necessary for emission, which is different from electromagnetism, which requires only a dipole.

[3]Note that the metric is more similar to the vector potential than to the electric and magnetic field in mathematical terms. As such, there are issues concerning gauge freedom, here diffeomorphism freedom, which need to be taken into account in actual calculations. They will be skipped over here.

The consequence of gravitational waves is that distances oscillate, since

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu = ds^2_{\text{flat}} + \Re\left(\epsilon_{\mu\nu}e^{i\eta_{\mu\nu}k^\mu x^\nu}\right)\epsilon_{\mu\nu}dx^\mu dx^\nu,$$

where ds^2_{flat} is the special relativity distance. Since the amplitude needs to be small, this cannot alter the causality structure. Hence, gravitational waves make themselves felt by changes in distances. This allows to detect them, e. g., by interferometry techniques. Making two waves interfere which travel at the same speed, but in different directions, their interference pattern will change if they are subjected to a gravitational waves, with an extend depending on the precise value of ϵ.

That can be realized in practice using lasers, and has been done so successfully. It is one of the most demanding applications from an experimental point of view. At the moment, earthbound detectors are sensitive to mergers of black holes and/or neutron stars. Future experiments will also be able to access other sources.

Chapter 8

Why there needs to be something beyond

At the end of 2009 the largest particle physics experiment so far has been started, the LHC at CERN. With proton-proton collisions at a center-of-mass energy of up to 14 TeV, possibly upgraded to 33 TeV by the end of the 2030ies, there are two major objectives. One was to complete the current picture of the standard model of particle physics by finding the Higgs boson. This has been accomplished, even though its properties remain still to be established with satisfactory precision, to make sure that it is indeed the Higgs boson predicted by the standard model.

The second objective is to search for new physics beyond the standard model. For various reasons it is believed that there will be new phenomena appearing in particle physics at a scale of 1 TeV, and thus within reach of the LHC. Though this is not guaranteed, there is motivation for it.

In fact, there are a number of reasons to believe that there exists physics beyond the standard model. These reasons can be categorized as being from within the standard model, by the existence of gravity, and by observations which do not fit into the standard model. Essentially all of the latter category are from astronomy, and there are currently very few observations in terrestrial experiments which are reproducible and do not satisfactorily agree with the standard model, and none which disagrees with sufficient statistical accuracy to claim observations of a deviation.

Of course, it should always be kept in mind that the standard model has never been completely solved. Though what has been solved, in particular using perturbation theory, agrees excellently with measurements, it is a highly non-linear theory. It cannot a-priori be excluded that some of the reasons to be listed here are actually completely within the standard model, once it is possible to solve it exactly.

Many of the observations to be listed can be explained easily, but not necessarily, by new physics at a scale of 1 TeV. However, it cannot be exclude that there is no new phenomena necessary for any of them up to a scale of 10^{15} GeV, or possibly up to the Planck scale of 10^{19} GeV, in which case the energy domain between the standard model and this scale is known as the great desert.

8.1 Inconsistencies of the standard model

There are a number of factual and perceived flaws of the standard model, which make it likely that it cannot be the ultimate theory.

The one most striking reason is the need for renormalization, as described in section 3.7: It is not possible to determine within the standard model processes at arbitrary high energies. The corresponding calculations break down eventually, and yield infinities. Though we have learned how to absorb this lack of knowledge in a few parameters, the renormalization constants, it is clear that there are things the theory cannot describe. Thus it seems plausible that at some energy scale these infinities are resolved by new processes, which are unknown so far. In this sense, the standard model is often referred to as a low-energy effective theory of the corresponding high-energy theory.

This theory can in fact also not be a (conventional) quantum field theory, as this flaw is a characteristic of such theories. Though theories exist which reduce the severity of the problem, supersymmetry at the forefront of them, it appears that it is not possible to completely resolve it, though this cannot be excluded. Thus, it is commonly believed that the high-energy theory is structurally different from the standard model, like a string theory. This is also the case when it comes to gravity, as discussed in the next section.

There are a number of aesthetic flaws of the standard model as well. First, there are about thirty different free parameters of the theory, varying by at least ten orders of magnitude and some of them are absurdly fine-tuned without any internal necessity. There is no possibility to understand their size or nature within the standard model, and this is unsatisfactory. Even if their origin alone could be understood, their relative size is a mystery as well. This is particularly true in case of the Higgs and the electroweak sector in general. There is no reason for the Higgs to have a mass which is small compared to the scale of the theory from which the standard model emerges. In particular, no symmetry protects the Higgs mass from large radiative corrections due to the underlying theory, which could make it much more massive, and therefore inconsistent with experimental data, than all the other standard model particles. Why this is not so is called the hierarchy problem, despite the fact that it could just be accidentally so, and not a flaw of the theory.

Even if this scale should be of the order of a few tens of TeV, there is still a factor of possibly 100 involved, which is not as dramatic as if the scale would be, say, 10^{15} GeV. Therefore, this case is also called the little hierarchy problem.

There is another strikingly odd thing with these parameters. The charges of the leptons and quarks need not be the same just because of QED - in contrast to the weak or strong charge, actually. They could differ, and in particular do not need to have the ratio of small integer numbers as they do - 1 to 2/3 or 1/3. This is due to the fact that the gauge group of QED is Abelian. However, if they would not match within the experimental precision of more than ten orders of magnitude, and also the number of charged particles under the strong and weak force were not perfectly balanced in each generation, then actually the standard model would not work, and neither would physics with neutral atoms. This is due to the development of a quantum anomaly, i. e., an effect solely related to quantizing the theory which would make it inconsistent, see section 6.11. Only with the generation structure of quarks and leptons with the assigned charges of the standard model, this can be avoided. This is ultimately mysterious, and no explanation exists for this in the standard model, except that it is necessary for it to work, which is unsatisfactory.

There is also absolutely no reason inside the standard model, why there should be more than one family, since the aforementioned cancellation works within each family independently and is complete. However, at least three families are necessary to have inside the standard model CP violating processes, i. e. processes which favor matter over anti-matter, see section 6.8. As discussed later, such processes are necessary for the observation that the world around us is made from matter. But there is no reason, why there should only be three families, and not four, five, or more. And if there should be a fourth family around, why would its neutrino be so heavy, more than 100 GeV, compared to the other ones, as can already be inferred from existing experimental data.

Another particle, however, could still be beyond the standard model: The Higgs boson. As the properties of the observed candidate are not yet beyond doubt, there is no certainty, however expected and likely, that it really is the Higgs. If it is not, the whole conceptual structure of the electroweak sector of the standard model immediately collapses, and is necessarily replaced by something else. Therefore, despite the fact that a relatively light Higgs boson is in perfect agreement with all available data, this is not necessarily the case in nature.

As an aside, there is also a very fundamental question concerning the Higgs sector. At the current time, it is not yet clear whether there can exist, even in the limited sense of a renormalizable quantum field theory, a meaningful theory of an interacting scalar field. This is the so-called triviality problem. So far, it is only essentially clear that the only consistent four-dimensional theory describing only a spin zero boson is one without any interactions. Whether this can be changed by adding additional fields, as in the standard model, is an open question. However, since this problem can be postponed to energy scales as high as 10^{15} GeV, or possibly even higher, in particular for a Higgs as light as the observed one, this question is not necessarily of practical relevance.

Finally, when extrapolating the running gauge couplings for a not-to-massive Higgs to an energy scale of about 10^{15} GeV, their values almost meet, suggesting that at this scale a kind of unification would be possible. However, they do not meet exactly, and this is somewhat puzzling as well. Why should this seem to appear almost, but not perfectly so?

8.2 Gravity

8.2.1 Problems with quantization

One obviously, and suspiciously, lacking element of the standard model is gravity. Up to now, no fully consistent quantization of gravity has been obtained. A problem is that a canonical quantized theory of gravity is perturbatively not renormalizable. This is visible when returning once more to its Lagrangian from from section 7.2,

$$\mathcal{L}_{\text{EH}} = \omega\left(\frac{1}{2\kappa}R - \frac{1}{\kappa}\Lambda + \mathcal{L}_M\right)$$

From this, it can be inferred that the coupling constant involved, κ or equivalently Newton's constant, is dimensionful, as is the cosmological constant Λ. Superficial (perturbative) calculations for (8.1) show that the theory is perturbatively non-renormalizable. As a consequence, an infinite hierarchy of independent parameters, all to be fixed by experiment, would be necessary to perform perturbative calculations, spoiling any predictivity of the theory. In pure gravity, these problems occur at NNLO, for matter coupled to gravity already at the leading order of radiative corrections. In particular, this implies that the theory is perturbatively not reliable beyond the scale $\sqrt{\kappa}$. Though this may be an artifact of perturbation theory, this has led to developments like supergravity based on local supersymmetry, loop quantum gravity, and string theory, to be discussed in sections 9.1.10, 9.5.2, and 9.5.3, respectively.

Irrespective of the details, the lack of gravity is an obvious flaw of the standard model. Along with this lack comes also a riddle. The natural quantum scale of gravity is given by the Planck scale

$$M_P = \frac{1}{\sqrt{G_N}} \approx 1.22 \times 10^{19}\ \text{GeV}.$$

This is 17 orders of magnitude larger than the natural scale of the electroweak interactions, and 19 orders of magnitude larger than the one of QCD. The origin of this mismatch is yet unsolved.

One of the more popular proposals how to explain this, discussed in section 9.5, is that this is only an apparent mismatch: The scales of gravity and the standard model are the same, but gravity is able to propagate also in additional dimensions not accessible by the standard model. The mismatch comes from the ratio of the total volumes, the bulk, and the apparent four dimensional volume, which is thus only a boundary, a so-called brane.

8.2.2 Asymptotic safety

Reiterating, the problem with the renormalizability of quantum gravity is a purely perturbative statement, since only perturbative arguments have been used to establish it. Thus, the possibility remains that the theory is not having such a problem, it is said to be asymptotically safe, and the problem is a mere artifact of perturbation theory. In this case, when performing a proper, non-perturbative calculation, no such problems would arise. In fact, this includes the possibility that κ imposes just an intrinsic cutoff of physics, and that this is simply the highest attainable energy, similarly as the speed of light is the maximum velocity. As a consequence, the divergences encountered in particle physics then only results from taking the improper limit $\kappa \to \infty$.

This or a similar concept of asymptotic safety can be illustrated by the use of a running coupling, this time the one of quantum gravity. The naive perturbative picture implies that the running gravitational coupling increases without bounds if the energy is increased, similarly to the case of QCD if the energy is decreased. Since the theory is non-linearly coupled, an increasing coupling will back-couple to itself, and therefore may limit its own growth, leading to a saturation at large energies, and thus becomes finite. This makes the theory then perfectly stable and well-behaved. However, such a non-linear back-coupling cannot be captured naturally by perturbation theory, which is a small-field expansion, and thus linear in nature. It thus fails in the same way as it fails at small energies for QCD. Non-perturbative methods, like renormalization-group methods or numerical simulations, have provided indications that indeed such a thing may happen in quantum gravity,.

As an aside, it has also been proposed that a similar solution may resolve both the hierarchy problem and the triviality problem of the Higgs sector of the standard model, when applied to the combination of Higgs self-coupling and the Yukawa couplings, and possibly the gauge couplings.

8.3 Observations from particle physics experiments

There are two generic types of particle physics experiments to search for physics beyond the standard model, both based on the direct interaction of elementary particles. One are those at very high energies, where the sheer violence of the interactions are expected to produce new particles, which can then be measured. The others are very precise (relatively) low-energy measurements, where very small deviations from the standard model are attempted to be detected. Neither of these methods has provided so far any statistically and systematically robust observation of a deviation from the standard model. Indeed, it happened quite often that a promising effect vanishes when the statistical accuracy is increased. Also, it has happened that certain effects have only been observed in some, but not all of conceptually similar experiments. In these cases, it can again be a statistical effect, or there is always the possibilities that some, at first glance, minor difference between the experiments can fake such an effect at one experiment, or can shadow it at another. So far, the experience was mostly that in such situation a signal was faked, but this then usually involves a very tedious and long search for the cause. Also, the calculation of the theoretically expected values is often involved, and it has happened that by improving the calculations, a previously observed mismatch is actually just an underestimation of the theoretical error.

At the time of writing, while new results are coming in frequently, there are very few remarkable results which should be mentioned, and which await further scrutiny. On the other hand, several hundreds of measurements have not yet indicated any deviations from known physics. One should then be especially wary: If enough measurements are performed, there is always a statistical probability that some of them show deviations, the so-called look-elsewhere effect, and a number of such deviations are expected.

One observation from precision experiments is that the magnetic moment of the muon is not quite the expected one, but shows a very small deviation. However, the theoretical computations are difficult, as virtual corrections from QCD enter, and it is therefore not clear whether this problem is just a theoretical one. Nonetheless, new experiments will be performed to at least obtain a better experimental value, while the theoretical calculations are continuously improved.

The next is that two experiments have in agreement reported results that cannot be explained with just the three ordinary neutrinos. However, these results are rather indirect, and at least another independent experiment will be required for confirmation.

So far, there are two remarkabley persistent deviations concerning muons. One is the deviation of the gyromagnetic factor of the muon from its expected value. The other is that its production in decays of mesons is different than expected. No deviations are yet seen for the electron and for the tauon. However, both are harder to deal with experimentally, either due to electron and hadronic background, respectively.

All of these observations are currently investigated further. Given the amount of statistics needed, it may take quite some time for a final answer about the reality of any of these deviations.

8.4 Astronomical observations

During the recent decades a number of cosmological observations have been made, which cannot be reconciled with the standard model of either particle physics or cosmology. These will be discussed here.

8.4.1 The cosmological standard model

To do requires knowledge of the current standard model of cosmology, describing our universe.

The basic theory to describe large-scale structures in the universe is general relativity introduced in chapter 7. There are three important observations:

- The universe looks in every spatial direction essentially the same on large scales. This is less well established as it may look like, and there are structure of hundred of millions of light-years across. Thus, while seeming plausible, the observational data put a big 'approximate' on this information.
- The further matter is away from us, the faster it recedes away from us, on average.
- The night sky is essentially dark.

Now the last item, while sounding innocent, is actually quite relevant. If there would be an infinite, everlasting universe, it could not be true. Because otherwise there would be enough stars, which had enough time, to fill the universe with a permanent (and actually infinite) glow. Thus, the universe must be dynamic.

Accepting the rough isotropy and homogeneity, the last option can then be explained with the existence of a big bang: The universe had a beginning. For this to be compatible with general relativity requires that a solution to (7.5) exists with the matter of the universe in it, and yielding a big bang.

This is indeed the case, the so-called Friedmann-Lemaître-Robertson-Walker universe. Spatial homogeneity and isotropy requires that, in a suitable coordinate system, the metric takes the Friedmann-Lemaitre-Robertson-Walker (FLRW) form

$$g_{\mu\nu} = -dt^2 + a(t)d^3\Sigma. \tag{8.1}$$

where $d^3\Sigma$ describes a homogeneous and isotropic spatial hypersurface. The only possibilities in general relativity for such a structure is

$$d^3\Sigma = \frac{dr^2}{1-kr^2} + s_k(r)^2 d^2\Omega = \frac{dr^2}{1-kr^2} + s_k(r)^2(d\theta^2 + \sin^2\theta d^2\phi),$$

where θ and ϕ are the usual polar and azimuthal angle. The quantity k determines the curvature of the hypersurface, yielding

$$\begin{aligned} s_{k>0} &= \frac{1}{\sqrt{k}}\sin\left(r\sqrt{k}\right) \\ s_{k=0} &= r \\ s_{k<0} &= \frac{1}{\sqrt{k}}\sinh\left(r\sqrt{-k}\right). \end{aligned}$$

Thus, positive values determine a spherical hypersurface, zero is a flat hypersurface, and a negative value is hyperbolically shaped. It is often customary to rescale $r \to rk^{-\frac{1}{2}}$, and thus distances are measured in units of $k^{-\frac{1}{2}}$. Then k becomes discrete with the three options ± 1 and 0. Note that throughout it will be assumed that the spatial hypersurface has trivial topology, i. e. no complicated boundary conditions, is simply connected and has no holes. This is compatible with observations, but it appears impossible in principle to constrain the topology to be trivial. However, a non-trivial topology would be discoverable if, e. g. the same objects are seen in different directions.

The factor a is determined by the solutions of Einstein's equation (7.5), and thus by the matter in the universe. Isotropy and homogeneity requires that the energy-momentum tensor becomes $T_{\mu\nu} = \mathrm{diag}(\rho, p, p, p)$, where $\rho(t)$ is the matter density and $p(t)$ is the pressure. They are not independent, but related by the equation of state of the matter. Not specifying the latter yet yields

$$\partial_t \rho = -3\frac{\partial_t a}{a}(\rho + p^2) \tag{8.2}$$

$$\frac{\partial_t^2 a}{a} = -\frac{4\pi G_N}{3}(\rho + 3p) + \frac{\Lambda}{3} \tag{8.3}$$

The value of k appears as constant of integration of equation (8.3),

$$\left(\frac{\partial_t a}{a}\right)^2 + \frac{k}{a^2} - \frac{\Lambda}{3} = \frac{8\pi G_N}{3}\rho,$$

and will thus be determined by the initial conditions of the density.

Note that replacing

$$
\begin{aligned}
\rho &\to \rho + \frac{\Lambda}{8\pi G_N} \\
p &\to p - \frac{\Lambda}{8\pi G}
\end{aligned}
$$

in (8.2-8.3) would eliminate the cosmological constant from Einstein's equation. This is equivalent to say that in this setting the cosmological constant behaves like matter satisfying the equation of state

$$0 < \rho_\Lambda = -p_\Lambda \tag{8.4}$$

i. e. like a positive matter density which exerts a negative pressure. Thus, the cosmological constant tends to blow up the universe.

To quantify the behavior of the universe the Hubble parameter is introduced as

$$H = \frac{\partial_t a}{a}. \tag{8.5}$$

Note that it is only spatially constant, but not in cosmological time t. It is usually quoted as its value today, where it has been evolved with a cosmological model when measured at a different time. It gives the rate of change of the universes spatial size normalized to its size. It follows that

$$H^2 = \frac{8\pi G}{3}\rho - \frac{k}{a}.$$

This can be used to define a critical density

$$\rho_c = \frac{3H^2}{8\pi G}$$

which implies that the quantity

$$\Omega = \frac{\rho}{\rho_c}$$

is, for k/a^2 negligible, a measure for the fate of the universe. If it is larger than 1, the normalized change of rate yields that the universe will eventually collapse. If it is equal one, it will reach an asymptotic size. And if it is smaller than one, the universe will expand forever. Current measurements strongly suggest $k \approx 0$ and $\Omega < 1$, and thus that the universe will expand forever.

What is finally needed to solve the system is the equation of state. Approximating the matter in the universe by a perfect fluid, the equation of state becomes

$$p = w\rho.$$

This implies that the cosmological constant behaves like a perfect fluid with $w = -1$. For matter with thermal energy substantially below the rest energy w is zero, while for ultrarelativistic matter, or massless particles, $w = 1/3$. Inserting these yields

$$a_\Lambda(t) \sim e^t \tag{8.6}$$
$$a_{\text{Matter}}(t) \sim t^{\frac{2}{3}} \tag{8.7}$$
$$a_{\text{Radiation}}(t) \sim t^{\frac{1}{2}} \tag{8.8}$$

To get a unified result, it is possible to combine different types of matter. This yields in terms of the fraction of the total density of each type in units of the critical density at a fixed time t_0

$$\frac{H^2}{H_0^2} = \frac{\Omega_{\text{Radiation}}}{a^4} + \frac{\Omega_{\text{Matter}}}{a^3} + \frac{\Omega_k}{a^2} + \Omega_\Lambda \tag{8.9}$$

where Ω_k is a suitable normalized version of k and H_0 is the value of H measured at the same fixed time t_0, usually today. This equation is, in principle, exactly solvable, given a very lengthy expression in terms of elliptic functions.

However, here just the important properties will be quoted for the observed values of the various Ω_i. The values are $\Omega_k \approx 0$, $\Omega_{\text{Matter}} \approx 0.25$, $\Omega_{\text{Radiation}} \approx 0.01$, and $\Omega_\Lambda \approx 0.74$, where dark matter is included in matter. These values are the ones measured today. This modelling of the universe is also known as the cold(-dark) matter scenario with cosmological constant, briefly ΛCDM. Cold, because the (dark) matter is non-relativistic.

This yields that the universe started out from $a = 0$, i. e. a point, the so-called big bang. At that point also time started, and all worldlines originate from this event. There is no notion, within general relativity, of a before or outside. That is a very crucial insight. However, at the same time, this is a singularity, and thus likely an indication for a breakdown of gravity and the need for quantum gravity. Afterwards, it started to expand spatially.

After the big bang the universe was extremely hot, and thus radiation dominated it, yielding an expansion like (8.8). With the matter becoming more and more diluted, the temperature eventually dropped, and the equation became that of matter, slowing down the expansion to (8.7). However, eventually the cosmological constant will take over, yielding an exponentially accelerated expansion of the universe. This happens roughly around now. Thus, the ultimate fate of the universe is to become infinitely large, provided that nothing yet unknown kicks in.

Note that this is the expansion of the spatial hypersurface of the universe itself. This does not mean that matter will be blown apart. Because other interactions, especially attractive gravitational and electric ones, exist, matter will not tend to become homogeneously distributed, but clusters. Thus, even in the exponential late-time acceleration cluster of matter will stay together on galactic scales, but likely not on larger scales.

While this scenario is working well with a lot of evidence , it is not fully able to explain all observations. Especially the age of the oldest stars and globular clusters, about 13 billion years, is older than the age of the universe as determined by (8.9), which is about 10 billion years. The currently most likely solution is that of an additional effect making the universe much bigger in its very early stage, the so-called inflation. What essentially has happened is that at time-scales somewhere between temperature of order the Planck energy, $M_P \sim 10^{19}$ GeV and the so-called GUT scale of 10^{15} GeV the universe expanded exponentially by a factor of about e^{60}. While such an effect can be mediated by any first order phase transition of the matter in the universe, there is no known matter which will do it at such short time scales and by this amount. The standard model of particle physics will only yield some factor of e^{4-5} and far too late to be compatible with observations. For the moment, this will be just accepted as a feature, which essentially modifies the initial condition for the solution of (8.9). It pushes the age of the universe to about 14 billion years.

One of the decisive assumption in these derivations was the homogeneity and isotropy of the universe. That is, of course, violated on small scales. For the derivations to make sense, this is no problem, as on sufficiently large scales it is (approximately) true. Large scales need to be scales which are still substantially smaller than the size of the universe. The later is estimated to be of order 100 billion lightyears today, so more than an order of magnitude larger than the visible universe, which is of order 10 billion lightyears, the distances traversable by light in the age of the universe. This is a consequence of inflation that the actual universe is larger than the visible one.

The largest observed structure in the universe, supercluster of galaxies as well as voids and filaments, regions of low densities of galaxies and one-dimensional structures with a high density of galaxies, are at least of order a few hundred million lightyears or more. Thus they approach a percent or more the size of the universe. With more observations incoming, this may worsen, eventually requiring to reproduce the evolution under less stringent assumptions on the structure of the universe. While this requires a numerical solution, anything where the fluctuations are not too extreme will still yield something approximately similar to the case presented here. Especially, any deviations can even influence the actual values of the Ω_i and H_0, as their definition is based on a homogeneous and isotropic universe.

8.4.2 Direct observations

Some observations not in line with known physics come from space-borne experiments, like the FERMI-Lat satellite and the AMS-2 spectrometer aboard the ISS. Both show an excess of anti-positrons over the expected yield in cosmic rays. If true, both would hint at an indirect detection of some new matter particles, perhaps dark matter, discussed in the next section. However, it can not yet be excluded that the rise in anti-positrons maybe due to not yet known or poorly understood conventional astrophysical sources.

Also, there are some hints from stellar evolutions which also may support indirect effects from additional matter particles, as some types of stars cool too efficiently as possible with just the known particles. But once again the effect may still be purely conventional.

Finally, every once in a while, some unexpected structure is observed in the spectrum of cosmic rays. Any such structure could indicate processes involving unknown particles. However, these structures so far usually faded away once the observational techniques have been improved or understanding of conventional astrophysical processes progressed.

8.4.3 Dark matter

One of the most striking observations is that the movement of galaxies, in particular how matter rotates around the center of galaxies, cannot be described just by the luminous matter seen in them and general relativity. That is actually a quite old problem, and known since the early 1930s. Also gravitational lensing, the properties of hot intergalactic clouds in galaxy clusters, the evolution of galaxy clusters and the properties of the large-scale structures in the universe all support this finding. In fact, most of the mass must be in the form of "invisible" dark matter. This matter is concentrated in the halo of galaxies, as analyses of the rotation curves and colliding galaxies show, as well as gravitational lensing. This matter cannot just be due to non-self-luminous objects like planets, brown dwarfs, cold matter clouds, or stellar black holes, as the necessary density of such objects would turn up in a cloaking of extragalactic light and of light from globular clusters. This matter is therefore not made out of any conventional objects, in particular, it is non-baryonic. Furthermore, it is gravitational but not electromagnetically active. It also shows different fluid dynamics (observed in the case of colliding galaxies) as ordinary luminous matter. Also, the dark matter cannot be strongly interacting, as it otherwise would turn up as bound in nuclei.

Thus this matter has to have particular properties. The only particle in the standard model which could have provided it would have been a massive neutrino. However, though the neutrinos do have mass, the upper limits on their mass is too low, and the flux of cosmic neutrinos too small, to make up even a considerable fraction of the dark matter. This can be seen by a simple estimate. If the neutrinos have mass and would fill the galaxy up to the maximum possible by Fermi-statistics, their density would be

$$n_\nu = \frac{p_F^3}{\pi^2}$$

with the Fermi momentum p_F in the non-relativistic case given by $m_\nu v_\nu$. Since neutrinos have to be bound gravitationally to the galaxy, their speed is linked via the Virial theorem to their potential energy

$$v_\nu^2 = \frac{G_N M_{\text{galaxy}}}{R},$$

with Newton's constant G_N and R the radius of the galaxy. Putting in the known numbers, and using furthermore that the observational results imply that n_ν, the total number of neutrinos approximated to be inside a sphere of the size of the galaxy, must give a total mass larger than the one of the galaxy leads to the bound

$$m_\nu > 100\text{eV}\left(\frac{0.001c}{3v_\nu}\right)^{\frac{1}{4}}\left(\frac{1\text{ kpc}}{R}\right)^{\frac{1}{2}},$$

yielding even for a neutrino at the speed of light a lower bound for the mass of about 3 eV, which is excluded by direct measurements in tritium decays.

Therefore, a different type of particles is necessary to fill this gap. In fact, many candidate theories for physics beyond the standard model offer candidates for such particles. But none has been detected so far, despite several dedicated experimental searches for dark matter. These experiments are enormously complicated by the problem of distinguishing the signal from background, in particular natural radioactivity and cosmic rays. The source of this matter stays therefore mysterious.

But not only the existence of dark matter, also its properties are surprising. The observations are best explained by dark matter which is in thermal equilibrium. But how this should be achieved if it is really so weakly interacting is unclear. This has lead to the idea that dark matter interacts with itself with a new force, which is stronger than just the gravitational interaction, but which does not couple to any standard model particles. In fact, it is also speculated that dark matter could consist out of many different particle species with many different forces.

On the other hand, the idea of gravitational bound dark matter is also problematic. In particular, there is no reason why it should neither form celestial dark bodies, which should be observable by passing in front of luminous matter by gravitational lensing, or why it should not be bound partly in the planets of our solar system. Only if its temperature is so high that binding is prohibited this would be in agreement, but then the question remains why it is so hot, and what is the origin of the enormous amount of energy stored in the dark matter.

It should be noted that there are also attempts to explain these observations by a departure of gravity from its classical behavior also at long distances. Though parametrizations exist of such a modification which are compatible with observational data, no clean explanation or necessity for such a modification in classical general relativity has been established. This proposal is also challenged by recent observations of colliding galaxies which show that the center-of-mass of the total matter and the center of luminous matter move differently, which precludes any simple modification of the laws of gravity, and is much more in-line with the existence of dark matter.

Another alternative is that the problem of asymptotic infrared safeness of quantum gravity may be related to the apparent existence of dark matter.

8.4.4 Inflation

A second problem is the apparent smoothness of the universe around us, while having at the same time small highly non-smooth patches, like galaxies, clusters, super clusters, walls and voids. In the standard model of cosmological evolution this can only be obtained by a rapid phase of expansion (by a factor $\sim e^{60}$) of the early universe, at temperatures much larger than the standard model scale, but much less than the gravity scale. None of the standard model physics can explain this, nor act as an agitator for it. In particular, it is also very complicated to find a model which at the same time explains the appearance of inflation and also its end.

However, the predictions of inflation have been very well confirmed by the investigation of the cosmic microwave background radiation, including non-trivial features and up to rather high precision.

8.4.5 Curvature, cosmic expansion, and dark energy

Another problem is the apparent flatness of the universe. Over large scales, the angle sum of a triangle is observed to be indeed π. This is obtained from the cosmic microwave background radiation, in particular the position of the quadrupole moment[1], but also that the large-scale structure in the universe could not have been formed in the observed way otherwise. For a universe, which is governed by Einstein's equation of general relativity,

[1]The homogeneity of the universe leads to a vanishing of the monopole moment and the dipole moment originates from the observer's relative speed to the background.

this can only occur if there is a certain amount of energy inside it. Even including the unknown dark matter, the amount of registered mass can provide at best about 30% of the required amount to be in agreement with this observation. The other contribution, amounting to about 70%, of what origin it may ever be, is called dark energy.

A second part of the puzzle is that the cosmic expansion is found to be accelerating. This is found from distant supernova data, which are only consistent if the universe expands accelerated today. In particular, other explanations are very hard to reconcile with the data, as it behaves non-monotonous with distance, in contrast to any kind of light-screening from any known physical process. Furthermore, the large-scale structures of the universe indicate this expansion, but also that the universe would be too young (about 10.000.000.000 years) for its oldest stars (about 12-13.000.000.000 years) if this would not be the case. For such a flat universe such an acceleration within the framework of general relativity requires a non-zero cosmological constant Λ, which appears in the Einstein equations (7.5). This constant could also provide the remaining 70% of the mass to close the universe, and is in fact a vacuum energy. Such a constant is covariantly conserved, since both $T_{\mu\nu}$ and the first two terms together are independently in general relativity, and thus indeed constant. However, the known (quantum) effects contributing to such a constant provide a much too large value for Λ, about 10^{40} times too large. These include quantities like the chiral and gluon condensates[2]. These are of order GeV, and in addition would have the wrong sign. What leads to the necessary enormous suppression is unclear.

8.4.6 Matter-antimatter asymmetry

In the standard model, matter and antimatter are not perfectly symmetric. Due to the CP violations of the electroweak force, matter is preferred over antimatter, i. e., decays produce more matter than antimatter, and also baryon and lepton number are not conserved quantities. However, this process is dominantly non-perturbative. The most striking evidence that this is a very weak effect is the half-life of the proton, which is larger than 10^{34} years. Indeed, only at very high-temperature can the effect become relevant.

After the big-bang, the produced very hot and dense matter was formed essentially from a system of rapidly decaying and recombining particles. When the system cooled down, the stable bound states remained in this process, leading first to stable nucleons and leptons in the baryogenesis, and afterwards to stable nuclei and atoms in the nucleosynthesis. Only over this time matter could have become dominant over antimatter, leading to the stable

[2]There are some subtleties involved here with renormalization in a quantum theory, and the true source of the problem is that the energy scale of gravity and the standard model are so vastly different. This is ignored here for the sake of the argument.

universe observed today. But the electroweak effects would not have been strong enough for the available time to produce the almost perfect asymmetry of matter vs. antimatter observed today, by a factor of about 10^{19}. Thus, a further mechanism must exist which ensures matter dominance today.

There is a profound connection to inflation. It can be shown that inflation would not have been efficient enough, if the number of baryons would have been conserved in the process. In particular, the almost-baryon-number conserving electroweak interactions would have permitted only an inflationary growth of e^{4-5} instead of e^{60}.

The possibility that this violation is sufficient to create pockets of matter at least as large as our horizon, but not on larger scales has been tested, and found to yield only pockets of matter much smaller than our horizon.

A further obstacle to a standard-model-conform breaking of matter-antimatter symmetry is the necessity for a first order phase transition. This is required since in a equilibrium (or almost equilibrium like at a higher-order transition), the equilibration of matter vs. anti-matter counters the necessary breaking. However, with a Higgs of about 125 GeV mass, this will not occur within the standard model.

8.5 How can new physics be discovered?

A task as complicated as the theoretical description of new physics is its experimental observation. One of the objectives of theoretical studies is therefore to provide signals on which the experiments can focus on. Here, some of the more popular experimental signatures, which can be calculated theoretically, will be introduced. It should be noted that one can differentiate between generic signatures and predictions. A generic signature is a qualitative deviation of an experiment from the prediction of known physics. A prediction is a quantitative statement about an experiment's outcome. The latter are experimentally simpler, as the experiment can be focused on a particular target, and the required sensitivity is known a-priori. The disadvantage is that a null result will only invalidate this particular prediction, and cannot make a statement about any other, differing prediction. Generic signatures require an experiment to check all possible results for anomalies, what is already for combinatorial reasons a serious obstacle, and in practice limited by computational resources and availability of personnel. But it has the advantage that no deviation from known physics could escape, in principle, up to the sensitivity the experiment can provide.

An important point with respect to sensitivity is the statistical significance of an observation[3]. Since the observed processes are inherently quantum, it is necessary to obtain a large sample (usually hundreds of millions of events) to identify something new, and still then an interesting effect may happening only once every hundredth million time. Experimentally identifying relevant events, using theoretical predictions, is highly complicated, and it is always necessary to quote the statistical accuracy (and likewise the systematic accuracy) for an event. Usually a three sigma (likeliness of 99.7%) effect is considered as evidence, and only five sigma (likeliness 99.9999%) are considered as a discovery, since in the past several times evidence turned in the end out to be just statistical fluctuations.

Especially, a single event can only in the rarest of cases be a reliable signal. Usually, background processes will create signatures similar to the searched-for events. Thus, the true signal is the amount of events more than just from any (known) background source, and one has to take into account also the signal-to-background ratio S/B. Usually, this ratio is much smaller than one, and thus background suppression is necessary. The way how to suppress background is usually process-dependent, and this is therefore beyond the current scope.

Besides statistical errors, there are the much harder to quantify systematic uncertainties, e. g. the energy resolution of an experiment. There is no way to guarantee that a systematic error has been correctly estimated. Thus in the past at several occasions a too optimistic assumption has created a fake signal, which went away once systematic effects have been correctly accounted for. These problems have to be kept in mind when interpreting experimental results.

8.5.1 Resonances

When in an interaction two particles exchange another particle, the cross-section of this process will be in lowest order proportional to the square of the propagator D of the exchanged particle, i. e.

$$\sigma \sim \frac{1}{(p^2 - M^2)^2},$$

where p is the energy transfer, and M is the mass of the new particle, see also (2.7). Therefore, the cross-section will exhibit a peak when the transferred energy equals the mass of the particle. Such resonances can be identified when the cross section is measured. If the mass does not belong to any known particle, this signals the observation of a new particle.

In practice, however, this simple picture is complicated by spin and other quantum numbers, interference, a finite decay width of the exchanged particles, and higher order

[3] Note that similar problems also apply to numerical simulations.

effects, and very often more than just the two original particles will appear in the final states. Identifying the peak in any particular channel of the interaction is therefore very complicated, and there are several instances of ghost peaks known, created by constructive interference. Still, this is one of the major ways of discovering a new particle directly. E. g., this has been how the Higgs particle has been discovered.

8.5.2 Missing energy

In principle akin to the concept of a resonance, the signature of missing energy is also associated directly with the (non-)observation of a particle. When two particles interact, the resulting particle may not be virtual, but real and stable, or at least sufficiently long-lived to escape the detector, in particular when its interactions with the standard model particles is small. Such a particle would surface in experiments as missing energy, i. e., the total observed energy would be smaller than before the collision, the remainder energy being carried away by the new particle. Looking for the smallest amount of missing energy would identify, using appropriate kinematics, the mass of the new particle. Dark matter, e. g., is assumed to produce precisely such a signature in collider experiments.

Again, in practice this concept is highly non-trivial, in particular due to muons, which can be compensated for a bit, and especially due to neutrinos. It is a highly complicated theoretically problem to identify further properties of missing energy events to identify cases where the missing energy can be unambiguously associated with the production of a new particle, even if this is as simple as an abundance of missing energy events.

8.5.3 Precision observables

A third possibility is the measurement of some quantities very precisely. Any deviation from the expected standard model value, if theoretically calculable with sufficient precision, is then indicating new physics. To identify the type and origin of such new physics, however, requires then careful theoretical calculations of all relevant models, and comparison with the measurement. Thus, a single deviation can usually only indicate the existence of new physics, but rarely unambiguously identify it. The advantage of such precision measurements is that they can usually be performed with much smaller experiments than collider experiments, but at the price of only very indirect information. Searches for dark matter or a neutron electric dipole moment larger than the standard model value are two examples of such low-energy precision experiments. But also collider experiments can perform precision measurements. Very important are electroweak oblique corrections, rather involved quantities determining electroweak radiative corrections to particle properties, or decay rates of mesons containing charm and, especially, bottom quarks.

8.5.4 Right-handed neutrinos

So far the leptons and neutrinos have been treated in the same way as the quarks. This consistent with all known results. However, there is a pecularity for neutrinos. Right-handed neutrinos are not charged under any of the gauge interactions, and only interact with the standard model by Yukawa interactions with the Higgs. As the size of the Yukawa couplings for the neutrinos is exceedingly small, there has not been any observation of right-handed neutrinos, in contrast for all other standard model particles. Hence, they are strictly speaking not necessary.

This leaves the question of how to then achieve the neutrino masses, which have been observed. The solution to this are so-called Majorana neutrinos. More appropriately would be to call them Weyl fermions, but in four dimensions this does not matter. In such a case only a two-dimensional Weyl spinor is necessary, i. e. only the left-handed component of the usual Dirac spinor is used. The original Dirac, or actually Majorana, spinor can be reconstructed as

$$\Psi = \begin{pmatrix} \psi^1 \\ \psi^2 \\ -\psi^{2*} \\ \psi^{1*} \end{pmatrix}.$$

The physical content is made manifest by performing a charge conjugation

$$C\Psi = \begin{pmatrix} 0 & i\sigma_2 \\ -i\sigma_2 & 0 \end{pmatrix} \begin{pmatrix} \psi \\ -i\sigma_2\psi^* \end{pmatrix}^* = \begin{pmatrix} \psi \\ -i\sigma_2\psi^* \end{pmatrix} = \Psi, \tag{8.10}$$

where C is an operation transforming a particle into its anti-particle. Thus, such a particle is its own anti-particle. A similar construction would not be possible for a quark. A scalar mass-term can be build from such a Weyl fermion as

$$m\psi^T(i\sigma_2)\psi$$

where σ_2 is the second Pauli matrix, and which is gauge-invariant. Note that in this case no Yukawa interaction of the neutrino with the Higgs is possible. The mass of such a Majorana neutrino has therefore a distinct other origin as those for Dirac neutrinos.

Because of the property (8.10), a propagating left-handed neutrino can convert into an anti-neutrino. This gives rise to new phenomena. An especially important signature is the so-called neutrinoless double β-decay. In this process, the anti-neutrino emitted during a β-decay of a nucleus converts to an neutrino, and allows then for second β-decay, in which the neutrino is absorbed by a neutron to emit a second electron and become a proton. Hence the name for the process. Similar phenomena can occur also in high-eenrgetic processes like the deacy of $W^\pm$ bosons, and thus yield also distinct signatures like same-sign lepton pairs in collider experiemnts. However, due to the feeble interaction of neutrinos it is kinematically complicated to show beyond doubt that not just a low-energy neutrino has escaped undetected.

Current experimental results remain consistent with both the existence and absence of a right-handed neutrino. It is currently not clear how long experiments will need to decide the matter.

8.6 Why one TeV?

There is the common expectation that something deviating from the standard model will show up at an energy scale of one to a few TeV.

One of the reason to expect this is that the Higgs plays a pivotal role to make the standard model consistent. If the Higgs should actually be only a particle (very) similar to the Higgs, this should show up at around 1 TeV. The reasons for this is given by the scattering cross-section of two W bosons to two W bosons, a process which in the standard model will occur a-plenty at that energy. At tree-level, the scattering amplitude with a non-standard-model Higgs is given by

$$M_{WW} = M_2(\cos\theta)\frac{s}{m_W^2} + M_1(\cos\theta)\ln\frac{s}{m_W} + M_0(\cos\theta),$$

where s is the center-of-mass energy, θ the angle between the scattered W bosons, and the amplitudes M_i describe the processes of scattering different polarizations of the W bosons. Unitarity, and thus preservation of causality, requires that this amplitude is bounded for $s \to \infty$, which is obviously not the case for the terms containing M_2 and M_1. Thus, for a center-of-mass energy significantly larger than the W boson mass m_W (in fact, about a TeV), unitarity is (naively) violated. In the standard model, interference with diagrams containing the Higgs remove this problem. But if it is not the standard-model Higgs, this should show up as severe deviations from the standard model at this energy scale.

A second reason to suspect something new is the energy scales so far encountered. In the standard model, so far the energy jumps between new particles is roughly an order of magnitude. There is no reason why this sequence should end now, so this lets it appear natural that something should happen at around 1 TeV.

Finally, most theories which can accommodate several of the shortcomings of the standard model would require to have very finely tuned parameters to be not visible at the

TeV scale. Though this is not impossible, there are very few examples known in nature where this occurs. This also motivates that new physics should be encountered at that scale. Especially, observations related to dark matter suggest that dark matter particles should have a mass in the TeV range, if they still react with the standard model.

Hence, all in all, though there is no guarantee that something interesting beyond a rather light Higgs has to happen at the TeV scale, there is quite some motivation in favor of it. Time will tell. And what this something may be, the following will try to give a glimpse of. But it should always be kept in mind that all of the reasons, and all of the following proposals can be completely incorrect. After all, nature is what it is, and there is no logically compelling reasons in favor (and quite a number against) of nature being to be as expected. This was the lesson of the turn of the 19th to the 20th century. Be prepared for everything, and keep in mind that at most one of the following proposals can be correct in the presented form.

Chapter 9

Candidate theories beyond the standard model

9.1 Supersymmetry

One of the most favored extensions of the standard model is supersymmetry, mainly due to its elegance and that there is a natural way how to construct a supersymmetric extension of the standard model.

9.1.1 A first encounter with supersymmetry

Supersymmetry is a symmetry which links the two basic types of particles one encounters in the standard model: bosons and fermions. In particular, it is possible to introduce operators which change a boson into a fermion and vice versa. Supersymmetric theories are then theories which are invariant under these transformations. This entails, of course, that such theories necessarily include both, bosons and fermions. In fact, it turns out that the same number of bosonic and fermionic degrees of freedom are necessary to obtain a supersymmetric theory. Hence to each particle there should exist a superpartner. Superpartners of bosons are usually denoted by the ending ino, so for a bosonic gluon the fermionic superpartner is called gluino. For the superpartners of fermions an s is put in front, so the super-partner of the fermionic top quark is the bosonic stop (and for quarks in general squarks).

Besides the conceptual interest in supersymmetric theories there is a number of phenomenological consequences which make these theories highly attractive. The most interesting of these is that quite a number of divergences, which have to be removed in ordinary quantum field theories by renormalization as discussed in section 3.7, drop out automatically.

A simple example for this concept is given by an infinite chain of harmonic oscillators. The Hamilton function for a harmonic oscillator is given, in terms of the bosonic creation and annihilation operators a and $a^\dagger$ fulfilling the algebra

$$\begin{aligned} \left[a, a^\dagger\right] &= aa^\dagger - a^\dagger a = 1 \\ [a, a] &= \left[a^\dagger, a^\dagger\right] = 0. \end{aligned} \tag{9.1}$$

by

$$H_B = a^\dagger a + \frac{1}{2}.$$

A non-interacting chain of these operators is therefore described by the sum

$$H_B = \sum_{n=1}^{N} \left(a^\dagger a + \frac{1}{2}\right). \tag{9.2}$$

If the number of chain elements N is sent to infinity, the contribution from the vacuum term diverges. Of course, in this simple example it would be possible to measure everything with respect to the vacuum energy, and just drop the diverging constant. However, imagine one adds a harmonic oscillator, which has fermionic quanta instead of bosonic ones. Such an oscillator is most easily obtained by replacing (9.2) with

$$H_F = \sum_{n=1}^{M} \left(b^\dagger b - \frac{1}{2}\right)$$

with the fermionic operators $b^\dagger$ and b satisfying the algebra

$$\begin{aligned} \left\{b, b^\dagger\right\} &= bb^\dagger + b^\dagger b = 1 \\ \{b, b\} &= \left\{b^\dagger, b^\dagger\right\} = 0. \end{aligned} \tag{9.3}$$

Such anti-commutation relations implement the Pauli principle: Since the creation operators square to zero, i. e. they are nil-potent, there is never more than one fermionic quanta.

For an infinite chain, the vacuum energy is negative. Adding both systems with the same number of degrees of freedom, N, the resulting total Hamiltonian

$$H = H_B + H_F = \sum_{n=1}^{N} \left(a^\dagger a + b^\dagger b\right) \tag{9.4}$$

has no longer a diverging vacuum energy. Actually, the Hamilton operator (9.4) is one of the simplest supersymmetric systems. Supersymmetry here means that if one exchanges bosons for fermions the Hamilton operator is left invariant. This will be discussed in detail below. The mechanism to cancel the vacuum energy is a consequence of supersymmetry, and this mechanism is one of the reasons which make supersymmetric theories very interesting.

From this cancellation it is also possible to obtain a 'natural' explanation why scalar particles should be of about the same mass as the other particles in a theory. There are some other reasons why supersymmetric theories are interesting:

- There exists a theorem that without supersymmetry it is not possible to embed the standard model (Lie groups) and general relativity (Poincare group) into a non-trivial (i.e. any other than a direct product) common structure (Coleman-Mandula theorem)
- Supersymmetric theories arise necessarily in the context of string theories
- Supersymmetry provides a simple possibility to extend the standard model
- Supersymmetry can be upgraded to include also gravity (called supergravity)
- Supersymmetry predicts relations between coupling constants which seem without relation without supersymmetry
- Supersymmetry provides such stringent constraints that many results can be shown exactly

In particular, it is widely expected that supersymmetry should be manifest at the 1-10 TeV scale, as this is the most likely range for the simplest extension of the standard model such that one ends up with the known particles with the right properties. However, there is no necessity for it.

Furthermore, there are also several reasons which make supersymmetry rather suspect. The most important one is that supersymmetry is not realized in nature. Otherwise the unambiguous predictions of supersymmetry would be that for every bosonic particle (e. g. the photon) a particle with the same mass, but different statistics (for the photon the spin-1/2 photino) should exist, which is not observed. The common explanation for this is that supersymmetry in nature has to be broken either explicitly or spontaneously. However, how such a breaking could proceed such that the known standard model emerges is not known. It is only possible to parametrize this breaking, yielding a enormous amount of free masses and coupling constants, for the standard model more than a hundred, while the standard model has only about thirty.

As a last remark of this introductory section it should be mentioned that supersymmetry offers also technical methods which can be used even if the underlying theory itself is not supersymmetric (which is done, e. g., in nuclear physics). So the use of supersymmetric concepts extends far beyond just possible extensions of the standard model.

9.1.2 Supersymmetric concepts

Supersymmetry is not a concept which requires either a field theory nor does it require special relativity. Its most simple form it takes in non-relativistic quantum mechanics. However, as fermions are involved, quantum effects are necessary. Such things as spin 1/2-objects, and thus fermions, cannot be described in the context of a classical theory.

9.1.2.1 Generators of supersymmetry

To be able to discuss a supersymmetric theory, it is of course necessary to have a Hilbert space which includes both bosons and fermions. One example would be the Hilbert space of the harmonic oscillator with only one element, the simplification of (9.4),

$$H = a^\dagger a + b^\dagger b. \tag{9.5}$$

The Hilbert space of this Hamilton operator is given by

$$|n_B n_F\rangle \tag{9.6}$$

where n_B is the number of bosons and n_F is the number of fermions. Note that no interaction between bosons and fermions occur. Such an interaction is not necessary to obtain supersymmetry. However, supersymmetry imposes very rigid constraints on any interactions appearing in a supersymmetric theory, as will be seen later. Furthermore, the number of degrees of freedom is the same: one bosonic and one fermionic oscillator.

Now, supersymmetry is based on a relation between fermions and bosons. Therefore, it will be necessary to introduce operators which can change a boson into a fermion. Such operators can be defined by their action on the states as

$$\begin{aligned} Q_+|n_B n_F\rangle &\sim |n_B - 1 n_F + 1\rangle \\ Q_-|n_B n_F\rangle &\sim |n_B + 1 n_F - 1\rangle. \end{aligned} \tag{9.7}$$

Note that since the maximum number of fermionic quanta is 1, Q_+ annihilates the state with one fermionic quanta, just as does the fermionic creation operator. Furthermore, the operator Q_- annihilates the fermionic vacuum, and Q_+ the bosonic vacuum, where vacuum is the state with zero quantas.

Now Q_+ and Q_- form symmetry transformation operators. However, for a theory to be supersymmetric, its Hamilton operator must be invariant under supersymmetry transformations, i. e.,

$$[H, Q_\pm] = 0. \tag{9.8}$$

To be able to check this, it is necessary to deduce the commutation relations of the bosonic and fermionic annihilation operators with $Q_\pm$. For this an explicit form of $Q_\pm$ in terms of a and b is necessary. Based on the action on the base state (9.7) it can be directly inferred that

$$\begin{aligned} Q_+ &= ab^\dagger \\ Q_- &= a^\dagger b. \end{aligned} \tag{9.9}$$

is adequate. From this also follows that $Q_\pm^\dagger = Q_\mp$, and both operators are adjoint to each other.

Testing now the condition (9.8) yields for Q_+

$$\left[a^\dagger a + b^\dagger b, ab^\dagger\right] = a^\dagger a a b^\dagger - ab^\dagger a^\dagger a + b^\dagger b a b^+ - ab^\dagger b^\dagger b. \tag{9.10}$$

Since two consecutive fermionic creation or annihilation operators always annihilate any state, it follows that

$$b^\dagger b^\dagger = bb = 0.$$

This property of operators is called nilpotency, the operators are nilpotent. This is a very important concept. However, currently this only implies that the last term in (9.10) is zero. Using the brackets (9.1) and (9.3), it then follows

$$b^\dagger a^\dagger a a - b^\dagger a - b^\dagger a^\dagger a a + ab^\dagger - ab^\dagger b^\dagger b = b^\dagger a - ab^\dagger = 0. \tag{9.11}$$

Here, once more nilpotency was used, and the fact that bosonic and fermionic creation and annihilation operators commute. The same can be repeated for Q_+. Hence the combined oscillator (9.5) is supersymmetric, and is called the supersymmetric oscillator. In fact this could be seen easier, as it is possible to rewrite (9.5) in terms of $Q_\pm$ as

$$\begin{aligned} H &= a^\dagger a + b^\dagger b + a^\dagger a b^\dagger b - b^\dagger b a^\dagger a \\ &= a a^\dagger b^\dagger b + a^\dagger a b b^\dagger \\ &= \left\{ab^\dagger, a^\dagger b\right\} = \{Q_+, Q_-\}. \end{aligned}$$

From this it follows directly that

$$\begin{aligned} [H, Q_+] &= [Q_+Q_-, Q_+] + [Q_-Q_+, Q_+] \\ &= Q_+Q_-Q_+ - Q_+Q_+Q_- + Q_-Q_+Q_+ - Q_+Q_-Q_+ = 0 \end{aligned}$$

This vanishes, as the operators $Q_\pm$ inherit the property of nilpotency directly from the fermionic operators. This can be repeated for Q_-, yielding the same result. As a consequence, any system which is described by the Hamilton operator $\{Q_+, Q_-\}$, where these can now be a more complex operator, is automatically supersymmetric.

The only now not manifest property is the hermiticity of the Hamilton operator. Defining the hermitian operators

$$\begin{aligned} Q_1 &= Q_+ + Q_- \\ Q_2 &= i(Q_- - Q_+) \end{aligned}$$

this can be remedied. By explicit calculation it follows that

$$H = Q_1^2 = Q_2^2. \tag{9.12}$$

Note that the hermitian operators Q_1 and Q_2 are no longer nilpotent. Also, since energy eigenstates are not changed by the Hamiltonian this implies that performing twice the supersymmetry transformation Q_1 or Q_2 will return the system to its original state. Finally, this implies that the supercharge of a system is an observable quantity.

Furthermore, it can be directly inferred that

$$\{Q_1, Q_2\} = 0.$$

It is then possible to formulate the supersymmetric algebra of the theory as

$$\begin{aligned} [H, Q_i] &= 0 \\ \{Q_i, Q_j\} &= 2H\delta_{ij}. \end{aligned} \tag{9.13}$$

This algebra is always available when the Hamilton operator takes the form (9.12). These algebras are often labeled with the number $\mathcal{N}$ of linearly independent generators of supersymmetric transformations. In the present case this number is $\mathcal{N} = 2$.

9.1.2.2 Spectrum of supersymmetric Hamilton operators

There are quite general consequences which can be deduced just from the algebra (9.13), which implies the form (9.12) of the Hamilton operator. Note that the only properties which have been used in the calculations previously of the supersymmetric operators were their nilpotency in the form of $Q_\pm$, and so they are not restricted to simple harmonic oscillators.

The first general consequence is that the spectrum is non-negative. This follows trivially from the fact that (9.12) can be written as the square of a hermitian operator, which has real eigenvalues. Given , e. g., that $|q\rangle$ is an eigenstate with eigenvalue q to the operator Q_1, it follows immediately that

$$H|q\rangle = Q_1^2|q\rangle = Q_1 q|q\rangle = q^2|q\rangle$$

and thus $|q\rangle$ is an energy eigenstate with energy q^2, which is positive or zero by the hermiticity of Q_1. Furthermore, let Q_2 act on $|q\rangle$, yielding

$$Q_2|q\rangle = |p\rangle.$$

This state is in general not an eigenstate of Q_1, although it is necessarily by virtue of (9.12) an eigenstate of Q_1^2. However, acting with Q_1 on $|p\rangle$ yields

$$Q_1|p\rangle = Q_1Q_2|q\rangle = -Q_2Q_1|q\rangle = -qQ_2|q\rangle = -q|p\rangle.$$

Here, it was assumed that q is not zero. Therefore, it is in fact an eigenstate, but to the eigenvalue $-q$, $|p\rangle = |-q\rangle$, so it is different from the original state $|q\rangle$. Therefore, this state is also an eigenstate of the Hamilton operator to the same eigenvalue. Hence, for all non-zero energies the spectrum of the Hamilton operator is doubly degenerate.

9.1.2.3 Breaking supersymmetry and the Witten index

If a symmetry is exact, its generator G not only commutes with the Hamilton operator

$$[H, G] = 0, \tag{9.14}$$

but the generator must also annihilate the vacuum, i. e., the vacuum $|0\rangle$ may not be charged with respect to the generator,

$$G|0\rangle = 0.$$

Otherwise, the commutator relation (9.14) would fail when applied to the vacuum[1]

$$[H, G]|0\rangle = HG|0\rangle - GH|0\rangle = HG|0\rangle.$$

If the vacuum would be invariant under G, the only other alternative to maintain the commutator relation, is that it also has to leave all other states invariant, and that it is a symmetry of the system. It then follows that supersymmetry is exact if

$$Q_1|0\rangle = 0 \text{ or } Q_2|0\rangle = 0.$$

[1]It is assumed that the energy is normalized such that the vacuum has energy zero.

This furthermore implies that if supersymmetry is unbroken the ground state has in fact energy zero. It will be shown below that this state has to be unique. It should be noted that there is a peculiarity associated with the fact that the Hamilton operator is just the square of the generator of the symmetry. Since if supersymmetry is broken

$$Q_i|0\rangle \neq 0 \Rightarrow Q_i^2|0\rangle \neq 0$$

the ground state can no longer have energy zero, but must have a non-zero energy. Therefore it is also doubly degenerate. Hence, the fact whether the ground state is degenerate or not is therefore indicative of whether supersymmetry is broken or not. This is described by the Witten index

$$\Delta = \mathrm{tr}(-1)^{N_F}.$$

N_F is the particle number operator of fermions. Therefore, this counts the difference in total number of bosonic and fermionic states

$$\begin{aligned}\Delta &= \sum \left(\langle n_F = 0|(-1)^{N_F}|n_F = 0\rangle + \langle n_F = 1|(-1)^{N_F}|n_F = 1\rangle\right)\\ &= n_B - n_F\end{aligned}$$

where n_B is the total number of bosonic states and n_F is the total number of fermionic states. All states appear doubly degenerate except at zero energy. Furthermore, these degenerate states are related by a supersymmetry transformation, and thus have differing fermion number. Hence, the Witten index is reduced to the difference in number of states at zero energy,

$$\Delta = n_B(0) - n_F(0).$$

It is therefore only nonzero if supersymmetry is intact, and zero if not.

9.1.2.4 Interactions and the superpotential

To obtain a non-trivial theory, it is necessary to modify the theory. To preserve the previous results, and obtain a supersymmetric theory, it will be necessary to keep the operator $Q_\pm$ nilpotent and fermionic. This can be obtained by replacing the definition (9.9) by

$$\begin{aligned}Q_+ &= Ab^\dagger\\ Q_- &= A^\dagger b,\end{aligned}$$

where the operators A and $A^\dagger$ are arbitrary bosonic functions of the original operators a and $a^\dagger$. In principle, they may also depend on bosonic quantities formed from b and

$b^\dagger$, like $b^\dagger b$. These would not spoil the nilpotency. For simplicity, this possibility will be ignored here.

In this case N_B is in general no longer a good quantum number, but N_F is trivially. So all states can be split into states with either $N_F = 0$ or $N_F = 1$, and can furthermore be labeled by their energy,

$$\begin{aligned} H|En_F\rangle &= E|En_F\rangle \\ N_F|En_F\rangle &= n_F|En_F\rangle. \end{aligned}$$

It is therefore useful to rewrite the states as

$$|En_F\rangle = \begin{pmatrix} |E0\rangle \\ |E1\rangle \end{pmatrix}.$$

In this basis, fermion operators act as matrices,

$$\begin{aligned} b^\dagger &= \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \\ b &= \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \end{aligned}$$

which act as, e. g.,

$$b^\dagger \begin{pmatrix} |x\rangle \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ |x\rangle \end{pmatrix}.$$

As a consequence the Hamilton operator takes the form

$$H = \begin{pmatrix} A^\dagger A & 0 \\ 0 & AA^\dagger \end{pmatrix} = \frac{1}{2}\{A, A^\dagger\} - \frac{1}{2}[A, A^\dagger]\sigma_3, \tag{9.15}$$

where σ_3 is the third of the Pauli matrices.

Non-trivial contributions in quantum theories usually only involve the coordinates, and not the momenta. The simplest way to obtain a non-trivial supersymmetric model is therefore to replace the definition of the bosonic creation and annihilation operators

$$\begin{aligned} a^\dagger &= \sqrt{\frac{m}{2}}\left(q - i\frac{p}{m}\right) \\ a &= \sqrt{\frac{m}{2}}\left(q + i\frac{p}{m}\right) \end{aligned}$$

with more general ones

$$\begin{aligned} A^\dagger &= \sqrt{\frac{1}{2}}\left(W(q) - i\frac{p}{\sqrt{m}}\right) \\ A &= \sqrt{\frac{1}{2}}\left(W(q) + i\frac{p}{\sqrt{m}}\right). \end{aligned}$$

The quantity $W(q)$ is called the super-potential, although it is not a potential in the strict sense, and its dimension is not that of energy.

Using the relation (9.15) it is directly possible to calculate the corresponding Hamilton operator. The anticommutator yields

$$\begin{aligned}\{A,A^\dagger\} &= \frac{1}{2}\left(\left(W-i\frac{p}{\sqrt{m}}\right)\left(W+i\frac{p}{\sqrt{m}}\right)+\left(W+i\frac{p}{\sqrt{m}}\right)\left(W-i\frac{p}{\sqrt{m}}\right)\right)\\ &= W^2+\frac{p^2}{m}.\end{aligned}$$

For the commutator the general relation

$$[p,F(q)]=-i\frac{dF}{dq}$$

and the antisymmetry of the commutator is sufficient to immediately read off

$$[A,A^\dagger]=i\left[\frac{p}{\sqrt{m}},W\right]=\frac{1}{\sqrt{m}}\frac{dW}{dq}.$$

The general supersymmetric Hamilton operator therefore takes the form

$$H=\frac{1}{2}\left(\frac{p^2}{m}+W^2\right)-\frac{\sigma_3}{2\sqrt{m}}\frac{dW}{dq}.$$

9.1.3 The simplest supersymmetric field theory

9.1.3.1 Setup

As discussed previously, it will be necessary to have the same number of fermionic and bosonic degrees of freedom. This requires at least two degrees of freedom, since it is not possible to construct a fermion with only one. Consequently, two scalar degrees of freedom are necessary. The simplest system with this number of degrees of freedom is a non-interacting system of a complex scalar field ϕ and a free Weyl fermion χ. A Weyl fermion is a two-component spinor, and loosely corresponds to a fermion which is identical to its anti-fermion. The corresponding Lagrangian is given by

$$\mathcal{L}=\partial^\mu\phi^\dagger\partial_\mu\phi+i\chi^\dagger\bar{\sigma}^\mu\partial_\mu\chi, \tag{9.16}$$

where the σ_μ are the Pauli matrices supplemented by the unit matrix, as in (6.27), and $\bar{\sigma}_\mu=(1,-\sigma_i)$.

The corresponding physics will be invariant under a supersymmetry transformation if the action is invariant[2]. Since it is assumed that the fields vanish at infinity this requires invariance of the Lagrangian under the supersymmetry transformation up to a total derivative.

[2]Actually, only up to anomalies. This possibility will be disregarded here, and is no problem for the cases presented.

The supersymmetry transformation must be of type

$$A' = A + \delta A.$$

It takes for the scalar field the form

$$\delta\phi = (-i\sigma_2\xi)^T\chi. \tag{9.17}$$

Herein, ξ is a constant, Grassmann-valued spinor. It thus anticommutes with χ^a. This is necessary in order to form a scalar, complex number from χ. By dimensional analysis, ξ has units of $1/\sqrt{\text{mass}}$. The corresponding transformation law for the spinor is

$$\delta\chi = \sigma_\mu\sigma_2\xi^*\partial_\mu\phi. \tag{9.18}$$

The pre-factor is fixed by the requirement that the Lagrangian is invariant under the transformation. The combination of ξ with σ_μ guarantees the correct transformation behavior of the expression under Lorentz transformation in spinor space. The derivative, which appears is necessary to construct a scalar under Lorentz transformation in space-time, and to obtain the correct mass-dimension. It is the only object which can be used for this purposes, as it is the only one which appears in the Lagrangian (9.16), besides the scalar field. The general structure is therefore fixed by the transformation properties under Lorentz transformation. That the pre-factors are in fact also correct can be shown by explicit calculation.

The set of fields ϕ and χ is called a super-multiplet. To be more precise, it is a left chiral super-multiplet, because the Weyl fermion has only one polarization of the helicity. It could be replaced with a Weyl fermion with the opposite helicity, yielding a right chiral multiplet, without changing the supersymmetry of theory, although, of course, the transformation is modified.

There should be a note of caution here. Unfortunately, this demonstration is insufficient to show supersymmetry of the quantized theory, and it will be necessary to modify the Lagrangian (9.16). This problem will become apparent when discussing the supersymmetry algebra. However, most of the calculations performed so far can be used unchanged.

The supersymmetry transformations (9.17) and (9.18) form an algebra, similar to the algebra (9.13) of the quantum mechanical case. This algebra can be used to systematically construct supermultiplets, and is useful for many other purposes. Therefore, this algebra will be constructed here, based first on the simplest examples of supersymmetry transformations (9.17) and (9.18) and will be generalized thereafter.

To construct the algebra, it is necessary to define (conserved) supercharges. These are given by

$$Q = \int d^3x\sigma^\nu\chi\partial_\nu\phi^\dagger,$$

and their conservation can be checked by explicit calculation.

Now it is possible to construct the algebra. First of all, the (anti-)commutators

$$[\bar{Q},\bar{Q}]=0 \qquad [Q,Q]=0 \qquad \{Q,Q\}=0 \qquad \{\bar{Q},\bar{Q}\}=0$$

all vanish, since in all cases all appearing fields (anti-)commute. There are thus, at first sight, one non-trivial commutator and one non-trivial anti-commutator. The commutator is given by

$$[\xi Q+\bar{\xi}\bar{Q},\eta Q+\bar{\eta}\bar{Q},\phi]=f^\mu(\eta,\xi)P_\mu, \tag{9.19}$$

with two parameters ξ and η and a fixed function f of them, and P_μ is the generator of translation, the momentum. In fact, this is not all, due to aforementioned subtlety involving the fermions. This will be postponed to later. The appearance of the momentum operator seems at first surprising. However, in quantum mechanics it has been seen that the super-charges are something like the square-root of the Hamilton operator, which is also applying, in the free case, to the momentum operator. This relation is hence less exotic than might be anticipated. Still, this implies that the supercharges are also something like the square-root of the momentum operator, which lead to the notion of the supercharges being translation operators in fermionic dimensions. This idea leads to the formalism of superspace, which will be skipped here. Another important side remark is that this result implies that making supersymmetry a local gauge symmetry this connection yields automatically a connection to general relativity, the so-called supergravity theories.

Hence, the algebra for the SUSY-charges will not only contain the charges themselves, but necessarily also the momentum operator. However, the relations are rather simple, as the supercharges do not depend on space-time and thus (anti-)commute with the momentum operator, as does the latter with itself.

Thus, the remaining item is the anticommutator of Q with $\bar{Q}$, given by

$$\{Q_a,Q_b^\dagger\}=(\sigma^\mu)_{ab}P_\mu. \tag{9.20}$$

This completes the algebra for the supercharges, which is in the field-theoretical case somewhat more complicated by the appearance of the generator of translations, but also much richer.

There are two remarks in order.

First, again, a subtlety will require to return to these results shortly.

Secondly, this results looks somewhat different from the quantum mechanical case (9.13), where there have been two supercharges, instead of one, which then did not anti-commute as is the case here. Also in field theories this may happen, if there are more than one supercharge. Since in the case of multiple supercharges there is still only one momentum operator, the corresponding superalgebras are coupled. In general, this requires the introduction of another factor δ^{AB}, where A and B count the supercharges. It appears in the (anti-)commutator of Q^A with $\bar{Q}^B$. The more drastic change is, however, that the anticommutator becomes modified to

$$\{Q_a^A, Q_b^B\} = \epsilon_{ab} Z^{AB},$$

and correspondingly for $\bar{Q}^A$ and $\bar{Q}^B$. The symbol ϵ_{ab} is anti-symmetric, and thus this anticommutator couples the algebra of different supercharges. Furthermore, the Z^{AB} are called the central charges, and are also anti-symmetric, but not an operator. The number of independent supercharges is labeled by $\mathcal{N}$. The theory with only one supercharge is thus an $\mathcal{N} = 1$ theory. In four dimensions, there are also $\mathcal{N} = 2$ and $\mathcal{N} = 4$ theories. The additional charges provide more constraints, so theories with more supercharges become easier to handle. However, currently there seems to exist no hint that any other than $\mathcal{N} = 1$-theories could be realized at energy scales accessible in the foreseeable future. Thus, here only this case will be treated.

9.1.3.2 Supermultiplets

A supermultiplet is a collection of fields which transform into each other under supersymmetry transformations. The naming convention is that a fermionic superpartner of a field a is called a-ino, and a bosonic superpartner s-a. Here, the structure of these supermultiplets will be analyzed.

One result of the algebra obtained in the previous subsection was that the momentum operator (anti-)commutes with all supercharges. Consequently, also P^2 (anti-)commutes with all supercharges,

$$[Q_a, P^2] = [\bar{Q}, P^2] = \{Q, P^2\} = \{\bar{Q}, P^2\} = 0.$$

Since the application of P^2 just yields the mass of a pure state, the masses of a particle s and its superpartner ss must be degenerate, symbolically

$$\begin{aligned} P^2|s\rangle &= & m^2|s\rangle \\ P^2|ss\rangle &= P^2Q|s\rangle = QP^2|s\rangle = Qm^2|s\rangle = m^2Q|s\rangle = & m^2|ss\rangle. \end{aligned}$$

This is very similar to the degenerate energy levels, which were encountered in the quantum mechanical case.

Further insight can be gained from considering the general pattern of the angular momentum in a supermultiplet. So far, it is only clear that in a supermultiplet bosons and fermions must be present, but not necessarily their spin. In the example, one was a spin-0 particle and one a spin-1/2 particle. This pattern of a difference in spin of 1/2 is general. In fact, the spin j in a supermultiplet can differ only by two units, and there are only two (times the number of internal quantum number) states in each supermultiplet. It does not specify the value of j, so it would be possible to have a supermultiplet with $j = 0$ and $j = 1/2$, as in the example above. This is called a chiral multiplet. Alternatively, it would be possible to have $j = 1/2$ and $j = 1$, the vectormultiplet, or $j = 2$ and $j = 3/2$, the gravity multiplet appearing in supergravity. However, any quantum field theory has the CPT-symmetry, thus for any supermultiplet also its CPT-transformed supermultiplet, and thus the corresponding antiparticles, have also to appear in the theory as well. Thus, there are always an even number of multiplets for charged states.

This is only true in the absence of central charges. In theories with more supercharges, the supermultiplets become larger. E.g. in $\mathcal{N} = 2$ SUSY, the supermultiplet contains two states with $j = 0$ and two with $m = \pm 1/2$. This is also the reason why theories with $\mathcal{N} \neq 1$ seem to be no good candidates for an extension of the standard model: SUSY, as will be seen below, requires that all superpartners transform the same under other transformations, like e. g. gauge transformations. In an $\mathcal{N} = 2$ theory, there would be a left-handed and a right-handed electron, both transforming under all symmetries in the same way. But in the standard model, the weak interactions couple differently to left-handed and right-handed electrons, and thus it is not compatible with $\mathcal{N} > 1$ SUSY. For $\mathcal{N} = 1$ SUSY, however, independent chiral multiplets (and, thus, more particles) can be introduced solving this problem. Since also the quarks (and the superpartners of the gauge-bosons therefore should also) are coupled differently in the weak interactions, this applies to all matter-fields, and thus it is not possible to enhance some particles with $\mathcal{N} = 1$ and others with $\mathcal{N} > 1$ SUSY.

9.1.3.3 Off-shell supersymmetry

Now, it turns out that the results so far are not complete, once the theory is quantized. In fact, the superalgebra is not closing off-shell, i. e. when due to quantum effects the equations of motions are not fulfilled: In a quantum theory exist virtual particles, i. e., particles which not only not fulfill the energy conservation, but also not their equations of motions. Hence such particles, which are called off (mass-)shell as discussed in section 2.3.1, are necessary. Thus the algebra so far is said to close only on-shell. Hence, although the theory described by the Lagrangian (9.16) is classically supersymmetric, is not so quantum-mechanically. Quantum effects break the supersymmetry of this model.

Hence, it is necessary to modify (9.16), to change the theory, to obtain one which is also supersymmetric on the quantum level. This can be done by the introduction of an auxiliary scalar field F, which has to be complex to provide two degrees of freedom, since off-shell a spinor has four, but not two, degrees of freedom. It is called auxiliary, as it has no consequence for the classical theory. The later can be most simply achieved by giving no kinetic term to this field. Thus, the modified Lagrangian takes the form

$$\mathcal{L} = \partial_\mu \phi^+ \partial^\mu \phi + \chi^\dagger i \bar{\sigma}^\mu \partial_\mu \chi + F^\dagger F. \tag{9.21}$$

It should be noted that this field has mass-dimension two, instead of one as the other scalar field ϕ. The equation of motions for this additional field is

$$\partial_\mu \frac{\delta \mathcal{L}}{\delta \partial_\mu F} - \frac{\delta \mathcal{L}}{\delta F} = F^\dagger = 0 (= F).$$

Thus, indeed, at the classical level it does not contribute.

Of course, if it should contribute at the quantum level, it cannot be invariant under a SUSY transformation. The simplest (and correct) guess is that this transformation should only be relevant off-shell. As it must make a connection to χ, the ansatz is

$$\begin{aligned} \delta_\xi F &= -i \xi^\dagger \bar{\sigma}^\mu \partial_\mu \chi \\ \delta_\xi F^\dagger &= i \partial_\mu \chi^\dagger \bar{\sigma}^\mu \xi, \end{aligned}$$

where ξ has been inserted as its transpose to obtain a scalar. The appearance of the derivative is also enforced to obtain a dimensionally consistent equation. This requires to also alter the other transformations to

$$\begin{aligned} \delta \chi &= \sigma_\mu \sigma_2 \xi^* \partial_\mu \phi + \xi F \\ \delta \chi^\dagger &= i \partial_\mu \phi^+ \xi^T (-i \sigma_2) \sigma^\mu + F^\dagger \xi^\dagger. \end{aligned}$$

Thus, without modifying the transformation law for the ϕ-field, the new Lagrangian (9.21) indeed describes a theory which is supersymmetric on-shell and off-shell.

9.1.4 Supersymmetric Abelian gauge theory

The simplest gauge theory to construct is a supersymmetric version of Maxwell theory. The photon is a boson with spin 1. Hence its superpartner, the photino, has to be a fermion. The photon has on-shell two degrees of freedom, so the photino has to be a Weyl fermion. Off-shell, however, the photon has three degrees of freedom, corresponding to the three different magnetic quantum numbers possible. So another bosonic degree of freedom is necessary to cancel all fermionic degrees of freedom. This other off-shell bosonic degree of freedom will be the so-called D field. This shows again the condition on the exact equality of the number of fermionic and bosonic degrees of freedom in a supersymmetric theory on-shell and off-shell.

Will this be a flavor of quantum electron dynamics then, just with the Dirac electron replaced by a Weyl one and one field added? The answer to this is a strict no. Since the supersymmetry transformation just acts on the statistical nature of particles it cannot change an uncharged photon into a charged photino. Thus, the photino has also to have zero charge, as does the D boson. The free supersymmetric Abelian gauge theory is thus a rather simple theory, as there are no interactions. Its form has thus to be of the type

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\lambda^\dagger\bar{\sigma}^\mu\partial_\mu\lambda + \frac{1}{2}D^2. \tag{9.22}$$

Note that the absence of charge also implies that no covariant derivative can appear. Hence, the photino λ has to be invariant under a gauge transformation, as has to be D. Thus the gauge dynamics is completely contained in the photon field.

To construct the SUSY transformation for A_μ, it is necessary that is has to be real, as A_μ is a real field. Furthermore, it has to be a Lorentz vector. Finally, since it is an infinitesimal transformation it has to be at most linear in the transformation parameter ξ. The simplest quantity fulfilling these properties is

$$\delta_\xi A_\mu = \xi^\dagger\bar{\sigma}_\mu\lambda + \lambda^\dagger\bar{\sigma}_\mu\xi.$$

As noted, the photino is a gauge scalar. It can therefore not be directly transformed with the field A_μ, but a gauge-invariant combination of A_μ is necessary. The simplest such quantity is $F_{\mu\nu}$. To absorb the two indices, and at the same time provide the correct transformation properties under Lorentz transformation, the SUSY transformations take the form

$$\begin{aligned}
\delta_\xi\lambda &= \frac{i}{2}\sigma^\mu\bar{\sigma}^\nu\xi F_{\mu\nu} + \xi D \\
\delta_\xi\lambda^\dagger &= -\frac{i}{2}\xi^\dagger\bar{\sigma}^\nu\sigma^\mu F_{\mu\nu} + \xi^\dagger D.
\end{aligned}$$

Finally, the SUSY transformation for the D-boson has to vanish on-shell, and thus should be proportional to the equations of motion of the other fields. Furthermore, it is a real field, and thus its transformation rule has to respect this, similarly as for the photon. The transformation rule such constructed is

$$\delta_\xi D = -i(\xi^\dagger\bar{\sigma}^\mu\partial_\mu\lambda - \partial_\mu\lambda^\dagger\bar{\sigma}^\mu\xi),$$

which has all of the required properties.

9.1.5 Supersymmetric Yang-Mills theory

A much more interesting theory will be the supersymmetric version of the non-Abelian gauge theory. The first thing to do is to count again degrees of freedom. The gauge field is in the adjoint representation. For SU(N) as a gauge group there are therefore N^2-1 independent gauge fields. On-shell, this has to be canceled by exactly the same number of fermionic degrees of freedom, so the same number of fermions, called gauginos. Quarks are, however, in the fundamental representation of the gauge group, and thus there are a different number, e. g. for SU(N) N. This cannot match, and the superpartners of the gauge fields, the gauginos, have therefore to be also in the adjoint representation of the gauge group. Therefore, they are completely different from ordinary matter fields like quarks or leptons. Hence, for each field in the standard model below it will be necessary to introduce an independent superpartner.

Of course, to close the SUSY algebra also off-shell, it will again be necessary to introduce another scalar. However, also this has to be in the adjoint representation of the gauge group. In this case, it is useful to write the Lagrangian explicitly in the index form. This yields the Lagrangian

$$\begin{aligned}
\mathcal{L} &= -\frac{1}{4}F_{\mu\nu}^a F_a^{\mu\nu} + i\lambda_a^\dagger\bar{\sigma}^\mu D_\mu^{ab}\lambda_b + \frac{1}{2}D^a D_a \\
F_{\mu\nu}^a &= \partial_\mu A_\nu^a - \partial_\nu A_\mu^a - ef^{abc}A_\mu^b A_\nu^c \\
D_\mu^{ab} &= \delta^{ab}\partial_\mu + ef^{abc}A_\mu^c,
\end{aligned} \tag{9.23}$$

where f^{abc} are the structure constants of the gauge group. D_μ^{ab} is the covariant derivative in the adjoint representation. The gauge bosons transform in the usual way, but the gaugino and the D-boson transform under gauge transformations in the adjoint representation. I. e., they transform like the gauge field, except without the inhomogeneous term as $g\lambda g^{-1}$, when written as algebra elements. Thus, even the D^2 term is gauge invariant, as no derivatives are involved.

The super transformations

$$\begin{aligned}
\delta_\xi A_\mu^a &= \xi^\dagger\bar{\sigma}_\mu\lambda^a + \lambda^{a\dagger}\bar{\sigma}_\mu\xi \\
\delta_\xi\lambda^a &= \frac{i}{2}\sigma^\mu\bar{\sigma}^\nu\xi F_{\mu\nu}^a + \xi D^a \\
\delta_\xi\lambda^{a\dagger} &= -\frac{i}{2}\xi^\dagger\bar{\sigma}^\nu\sigma^\mu F_{\mu\nu}^a + \xi^\dagger D^a \\
\delta_\xi D^a &= -i(\xi^\dagger\bar{\sigma}^\mu D_\mu^{ab}\lambda^b - D_\mu^{ab}\lambda^\dagger\bar{\sigma}^\mu\xi)
\end{aligned}$$

as the most simple generalization of the Abelian case are correct.

9.1.6 Supersymmetric QED

The Abelian gauge theory contained only an uncharged fermion, the photino. To obtain a supersymmetric version of QED a U(1)-charged fermion is necessary. Since this cannot be introduced into the vector supermultiplet of the photon, the most direct way to introduce it is by the addition of a chiral supermultiplet which will be coupled covariantly to the vector supermultiplet. Of course, this introduces only a Majorana electron, but this will be sufficient for the beginning.

The minimally coupled version is

$$\mathcal{L} = (D_\mu\phi)^\dagger D^\mu\phi + i\chi^\dagger\bar{\sigma}^\mu D_\mu\chi + F^\dagger F - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\lambda^\dagger\bar{\sigma}^\mu\partial_\mu\lambda + \frac{1}{2}D^2. \tag{9.24}$$

This theory contains now the photon, its superpartner the photino, the (Majorana) electron, its superpartner the selectron, and the auxiliary bosons F and D. In essence, this Lagrangian is the sum of the non-interacting Wess-Zumino model and the supersymmetric Abelian gauge theory, where in the former the derivatives have been replaced by gauge covariant derivatives. This implies that all newly added fields, ϕ, χ, and F, are charged, and transform under gauge transformations. Only the photino and the D-boson remain uncharged. The supersymmetry transformations have to be somewhat generalized, of course, to include the gauge symmetry correctly.

The Lagrangian (9.24) is indeed, in contrast to ordinary QED, not the most general[3] one which can be written with this field content and which is supersymmetric, gauge-invariant, and renormalizable. It is possible to add additional interactions which do have all of these properties. However, this will require also some minor modifications of the supersymmetry transformations.

In fact, it is possible to add further interactions between the matter and the gauge sector to (9.24). These have to be, of course, gauge-invariant, and will also be chosen to be renormalizable. The final Lagrangian for supersymmetric QED, requiring again slightly augmented supersymmetry transformations, reads

$$\begin{aligned}\mathcal{L} &= (D_\mu\phi)^\dagger D^\mu\phi + i\chi^\dagger\bar{\sigma}^\mu D_\mu\chi + F^\dagger F - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\lambda^\dagger\bar{\sigma}^\mu\partial_\mu\lambda + \frac{1}{2}D^2\\ &\quad -\sqrt{2}q((\phi^\dagger\chi)\lambda + \lambda^\dagger(\chi^\dagger\phi)) - q\phi^\dagger\phi D.\end{aligned}$$

A number of remarks are in order. First, though two essentially independent sectors have been coupled, this does not lead to a product structure with two independent supersymmetry transformations, but only to one common transformation. The deeper reason for this

[3] Here, and in the following, Wess-Zumino-like couplings, which are in fact gauge-invariant and supersymmetric, will be ignored.

is the appearance of the momentum operator in both supersymmetry algebras, inevitably coupling both. The second is that the combination of gauge symmetry and supersymmetry required the introduction of interaction terms between both sectors to yield a supersymmetric theory. This interaction is so strongly constrained that its structure is essentially unique, and besides no new coupling constants appear compared to the two original theories. Again, supersymmetry is tightly constraining. Third, the equation of motion for the D-boson yields $D = q\phi^\dagger\phi$. As a consequence, integrating out the D-boson lets a quadratic interaction term $-q^2/2(\phi^\dagger\phi)^2$ appear. This is a necessary ingredient for the possibility of a Higgs effect, driven by a selectron condensation. Thus, even the simplest non-trivial supersymmetric gauge-theory provides much more possibilities than ordinary QED.

9.1.7 Supersymmetric QCD

Again, having the standard model in mind, it is necessary to generalize supersymmetric QED to a non-Abelian version, the simplest of which is supersymmetric QCD. However, in the following the gauge group will not be made explicit, and thus the results are valid for any (semi-)simple Lie group as gauge group.

Since in QCD, and in the standard model in general, the fermions are in the fundamental representation, while the gauge fields are in the adjoint representation, it is not possible to promote somehow the matter fields to the super partners of the gauge bosons, despite these being charged in the supersymmetric version of Yang-Mills theory. It is therefore again necessary to introduce the matter fields as independent fields, together with their superpartners, and couple them minimally to the gauge fields. Therefore, besides the gauge-fields, the gluons, their super-partner, the gluinos, the D-bosons, there will be the quarks, their superpartners, the squarks, and the F-bosons.

Fortunately, the results of supersymmetric QED, together with those for supersymmetric Yang-Mills theory, can be generalized immediately. This yields

$$\begin{aligned}
\mathcal{L} &= -\frac{1}{4}F^\alpha_{\mu\nu}F^\alpha_{\mu\nu} + i\lambda^{\alpha+}\bar{\sigma}^\mu D^{\alpha\beta}_\mu\lambda^\beta + \frac{1}{2}D^\alpha D^\alpha \\
&\quad + D^{ij}_\mu\phi_j D^\mu_{ik}\phi^\dagger_k + \chi^\dagger_i i\bar{\sigma}^\mu D^{ij}_\mu\chi_j + F^\dagger_i F_i \\
&\quad + M\phi_i F_i - \frac{1}{2}M\chi_i\chi_i + \frac{1}{2}y_{ijk}\phi_i\phi_j F_k - \frac{1}{2}y_{ijk}\phi_i\chi_j\chi_k + h.c. \\
&\quad - \sqrt{2}e(\phi^\dagger\chi\lambda + \lambda^\dagger\chi^\dagger\phi) - g\phi^\dagger\phi D,
\end{aligned}$$

where it should be noted that inter-representation couplings equal, e. g.,

$$\begin{aligned}
\phi^\dagger\chi\lambda &= \phi^\dagger\tau^\alpha\chi\lambda^\alpha = \phi^\dagger_i\tau^\alpha_{ij}\chi^a_j\lambda^\alpha_a \\
\phi^\dagger\phi D &= \phi^\dagger\tau^\alpha\phi D^\alpha = \phi^\dagger_i\tau^\alpha_{ij}\phi_j D^\alpha.
\end{aligned}$$

Note that the coupling constants y acquired gauge indices in the fundamental representation. Similar to the Higgs-fermion coupling in the standard model, relations between the different elements of y ensure gauge invariance of these couplings, and these could be determined by explicit evaluation. Note that the interaction between the matter sector in the fundamental representation and the gauge-sector in the adjoint representation is completely fixed by gauge symmetry and supersymmetry, and there is no room left for any other interactions. In particular, despite the fact that six fields interact with altogether 11 interaction vertices, there are only three independent parameters, the mass parameter M, the Yukawa coupling y, which is constrained by gauge symmetry, and the gauge-coupling e. Note that the mixed term appearing from the supersymmetry transformation of F coupling in the Yukawa term vanishes due to the antisymmetry of the y-matrix, which is necessary to ensure gauge invariance.

It is worthwhile to evaluate the terms including a D explicitly after using its equation of motion, which reads

$$\frac{\delta \mathcal{L}}{\delta D_\alpha^\dagger} = D^\alpha - e\phi^\dagger \tau^\alpha \phi = 0.$$

This yields the total self-interaction (or potential V) of the ϕ field, after integrating out F as

$$V = |M\phi^\dagger \phi|^2 + \frac{1}{2}e^2(\phi^\dagger \tau^\alpha \phi)^2 - y_{ijk} y^*_{lmk} \phi_i \phi_j \phi_i \phi_j \phi_l^\dagger \phi_m^\dagger.$$

This potential is positive definite, and all of the couplings are uniquely defined. Thus, in contrast to the case of the standard model, the Higgs-potential, as this is the role the squark plays, is completely determined due to supersymmetry. This puts, at least perturbatively, strong constraints on the Higgs mass in the supersymmetric version of the standard model.

9.1.8 Supersymmetry breaking

All of the theories explicitly discussed so far have manifest supersymmetry. As a consequence, as it was shown generally, the masses of the particles and the sparticles have to be degenerate. This is not what is observed in nature. It is therefore necessary to break supersymmetry in some way.

At the present time, no satisfactory mechanism exists for the spontaneous breaking of supersymmetry for supersymmetric theories compatible with the observations. Therefore, it is necessary to parametrize this lack of knowledge in the form of explicitly supersymmetry breaking terms[4]. However, if some of the specific properties of supersymmetric theories,

[4] Of course, it cannot be excluded that supersymmetry in nature is in fact explicitly broken.

in particular the better renormalization properties and the protection of the scalar masses, should be retained, it is not possible to add arbitrary terms to the Lagrangian.

To ensure the survival, or at least only minor modification, of these properties, it is necessary to restrict the explicit supersymmetry breaking contributions to soft contributions. This ensures that supersymmetry becomes effectively restored at large energies. To ensure such a property, it is necessary that the terms contain coupling constants which have a positive energy dimensions. In case of the theories discussed so far, these can appear in the form of two types of terms.

One type of such terms are clearly masses which do not emerge from a superpotential. In particular, such terms are allowed even without integrating out the F-bosons. Such masses are possible for both, the bosonic and the fermionic fields, but not for the auxiliary fields, as their mass dimensions are not permitting renormalizable mass terms. However, these cannot appear for gauge bosons, as in standard gauge theories. Thus, e. g., in the supersymmetric QCD, such terms would be of type

$$-\frac{1}{2}\left(m_\lambda\lambda\lambda + m_\chi\chi\chi + m_\lambda^\dagger\lambda^\dagger\lambda^\dagger + m_\chi^\dagger\chi^\dagger\chi^\dagger + 2m_\phi^2\phi^\dagger\phi\right)$$

Note that all mass terms are gauge-invariant. Also, since this is an effective parametrization, all masses, m_λ, m_χ, and m_ϕ, are independent parameters. Furthermore, these masses may be complex. When the F-field would be integrated out, the additional mass terms for the electrons and the selectrons would mix with those introduced above. Not only is such a Lagrangian not explicitly supersymmetric anymore, but, since the superpartners are no longer mass-degenerate, also the spectrum is manifestly no longer supersymmetric.

Furthermore, an additional possibility are three-boson couplings. These couplings would be of type

$$a_{ijk}\phi_i\phi_j\phi_k + b_{ijk}\phi_i^\dagger\phi_j\phi_k + c_{ijk}\phi_i^\dagger\phi_j^\dagger\phi_k + d_{ijk}\phi_i^\dagger\phi_j^\dagger\phi_k^\dagger,$$

where, of course, not all couplings are independent, but are constrained by gauge-invariance and hermiticity. That such couplings break supersymmetry is evident, as they not include the particles and their superpartners equally.

All in all, to the three independent parameters of supersymmetric QCD, the gauge coupling, and the two Wess-Zumino parameters g and M, four more have emerged, three masses, and at least one three-boson coupling. All of these additions are in fact supersymmetry breaking. Inside such a model, there is no possibility to predict the values of the additional coupling constants. However, in cases where the (broken) supersymmetric theory is just the low-energy limit of a unified theory at some higher scale, this underlying theory can provide such predictions, at least partially. E. g., in the cases above, the expectation values of fields have been obtained in terms of the masses and the coupling constants of the fields, or vice versa.

This already closes the list of generally possible soft supersymmetry breaking terms, although for specific models much more possibilities may exist.

9.1.9 The minimal supersymmetric standard model

From the previous examples of simple theories it is clear that a supersymmetric transformation cannot change any quantum numbers of a field other than the spin. In the standard model, however, none of the bosons have the same quantum numbers as any of the fermions. Hence, to obtain a supersymmetric version of the standard model, it will be necessary to construct for each particle in the standard model a new superpartner. Of course, additional fields need also be included, like appropriate F-bosons and D-bosons. Here, the supersymmetric version of the standard model with the least number of additional fields will be introduced, the so-called minimal supersymmetric standard model (MSSM). This requires the following new particles

- The photon is uncharged. Therefore, its superpartner has also to be uncharged, and to be a Weyl fermion, to provide two degrees of freedom. It is called the photino

- The eight gluons carry adjoint color charges. This requires eight Weyl fermions carrying adjoint color charges as well, which are called gluinos

- Though massive, the same is true for the weak bosons, leading to the charged superpartners of the W-bosons, the wino, and of the Z-boson, the zino, together the binos. Except for the photino all gauginos, the gluinos and the binos, interact through covariant derivatives with the original gauge-fields

- Assuming that all neutrinos are massive, and not just the minimum number of two currently required by experiment, no distinction between left- and right-handed fermions would be necessary, but the weak interaction violates still parity, requiring a differentiation. However, no new interactions should be introduced due to the superpartners, and hence the superpartners cannot be spin-1 bosons, which would be needed to be gauged. Therefore these superpartners are taken to be scalars, called the sneutrinos, the selectron, the smuon, and the stau

- The same applies to quarks, requiring the fundamentally charged squarks

- The Higgs requires a fermionic superpartner, the higgsino. However, requiring supersymmetry forbids that the Higgs has the same type of coupling to all the standard model fields. Therefore, a second set of Higgs particles is necessary, with their corresponding superpartners. These are the only new particles required

- Of course, a plethora of auxiliary D and F bosons will be necessary

It is necessary to discuss the formulation of these fields, and the resulting Lagrangian, in more in detail.

The electroweak interaction acts differently on left- and right-handed fermions. It thus fits naturally to use independent left- and right-handed chiral multiplets to represent the leptons and the quarks, together with their bosonic superpartners, the squarks and sleptons. However, both Weyl-spinors can be combined into a single Dirac-spinor, as the total number of degrees of freedom match. The electroweak interaction can then couple by means of the usual $1 \pm \gamma_5$ coupling asymmetrically to both components. In this context, it is often useful that a charge-conjugated right-handed spinor is equivalent to a left-handed one, permitting to use exclusively left-handed chiral multiplets, where appropriate charge-conjugations are included.

Furthermore, for each flavor it is necessary to introduced two chiral multiplets, giving a total of 12 supermultiplets for the quarks and the leptons each. The mixing of quark and lepton flavors proceeds in the same way as in the standard model. This requires thus already the same number of parameters for two CKM-matrices as in the standard model.

The gauge-bosons are much simpler to introduce. Again, a gauge-multiplet is needed for the eight gluons. As electroweak symmetry breaking is not yet implemented, and actually cannot without breaking supersymmetry as well as noted below, there is actually a SU(2) gauge multiplet and an U(1) gauge multiplet, which do not yet represent the weak gauge bosons and the photon. Instead, before mixing, these are called $W^{\pm}$ and W^0, and B. Only after mixing, the W^0 and the B combine to the Z and the conventional photon.

Concerning interactions, there are, of course, the three independent gauge couplings of the strong, weak, and electromagnetic forces, to be denoted by g, g' and e. After electroweak symmetry breaking, g' and e will mix, just as in the standard model. Note that the coupling of the gauge multiplets and the chiral multiplets will induce additional interactions, as in case of supersymmetric QED and QCD. However, no additional parameters are introduced by this.

It remains to choose the superpotential for the 24 chiral multiplets, and to introduce the Higgs field. The parity-violating weak interactions imply that a mass-term, the component of the superpotential proportional to $\chi\chi$, is not gauge-invariant, since a product of two Weyl-spinors of the type $\chi\chi$ is not, if only the left-hand-type component is transformed under such a gauge transformation. This is just as in the standard model. Therefore, it is not possible, also in the MSSM, to introduce masses for the fermions by a tree-level term. Again, the only possibility will be one generated dynamically by the coupling to the Higgs field[5]. There, however, a problem occurs. In the standard model this is mediated by a Yukawa-type coupling of the Higgs field to the fermions. However, it is necessary, for the sake of gauge-invariance, to couple the two weak charge states of the fermions differently, one to the Higgs field, and one to its complex conjugate. This is not possible in a supersymmetric theory, as the superpotential can only depend on the field, and not its complex conjugate. It is therefore necessary to couple both weak charge states to different Higgs fields. This makes it possible to provide a gauge-invariant Yukawa-coupling for both states, but requires that there are two instead of one complex Higgs doublets in the MSSM. Furthermore, also these fields need to be part of a chiral supermultiplet, and therefore their partners, the higgsinos, are introduced as Weyl fermions. However, for the Higgs bosons, which have no chirality, the limitations on a mass-term do not apply, and one can therefore be included in the superpotential.

In principle, it would also be possible to add contributions which couple squarks and/or sleptons to quarks and/or leptons. Since these carry the same charges as their counterparts, they will also carry the same lepton and baryon numbers. As a consequence, the squarks $\tilde{Q}$ from the multiplets including the particle-like left-handed quarks carry baryon number $B = 1/3$, while those from the anti-particle-like right-handed chiral multiplets carry baryon number $-1/3$. Similarly, the fields $\tilde{L}$ carry lepton number $L = 1$ and $\bar{L}$ -1. Thus such interactions violate lepton number conservation and baryon number conservation. On principal grounds, there is nothing wrong with this, as both quantities are violated by non-perturbative effects also in the standard model. However, these violations are very tiny, well below nowadays experimental detection limit. But additional such interactions, on the other hand, would provide very strong, and experimentally excluded, violations of both numbers, as long as the coupling constants would not be tuned to extremely small values. Such a fine-tuning is undesirable.

However, such direct terms could be excluded, if all particles in the MSSM would carry an additional multiplicatively conserved quantity, called R-parity, which is defined as

$$R = (-1)^{3B+L+2s},$$

with s the spin. Such a quantum number would be violated by the offensive interactions, and these are therefore forbidden[6]. The contribution $2s$ in the definition of R implies

[5]Actually, a gaugino condensation would also be possible, if a way would be known how to trigger it and reconcile the result with the known phenomenology of the standard model.

[6]A non-perturbative violation of R-parity notwithstanding, as for baryon and lepton number in the standard model.

that particles and their superpartners always carry opposite R-parity. This has some profound consequences. One is that the lightest superparticle cannot decay into ordinary particles. It is thus stable. This is actually an unexpected bonus: If this particle would be electromagnetically uncharged, it is a natural dark matter candidate. However, it has also to be uncharged with respect to strong interactions, as otherwise it would be bound in nuclei. This is not observed, at least as long as its mass is not extremely high, which would be undesirable, as then all superparticles would be very massive, preventing a solution of the naturalness problem by supersymmetry. Therefore, it interacts only very weakly, and is thus hard to detect. Not surprisingly, it has not been detected so far, if it exists.

This concludes the list of additional particles and interactions required for the MSSM. As discussed previously, unbroken supersymmetry requires the particles and their superpartners to have the same mass. This implies that the superpartners should have been observed if this is the case, as the selectron should also have only a mass of about 511 keV, as the electron, and must couple in the same way to electromagnetic interactions. Therefore, it would have to be produced at the same rate in, e. g., electromagnetic pion decays as electrons. This is not the case, and therefore supersymmetry must be (completely) broken. Explicit searches in particle physics experiments actually imply that the superpartners are much heavier than their partners of the standard model, as non has been discovered so far. Hence, supersymmetry is necessarily strongly broken in nature.

Since no mechanism is yet known how to generate supersymmetry breaking in a way which would be in a accordance with all observations, and without internal contradictions, this breaking is only parametrized in the MSSM, and requires a large number of additional free parameters. Together with the original about 30 parameters of the standard model, these are then more than a 130 free parameters in the MSSM. Exactly, there are 105 additional parameters in the MSSM, not counting contributions from massive neutrinos.

These parameters include masses for all the non-standard model superpartners, that is squarks and sleptons, as well as photinos, gluinos, winos, and binos. Only for the Higgsinos it is not possible to construct a gauge-invariant additional mass-term, due to the chirality of the weak interactions, in much the same way as for quarks and leptons in the standard model. One advantage is, however, offered by the introduction of these free mass parameters: It is possible to introduce a negative mass for the Higgs particles, reinstantiating the same way to describe the breaking of electroweak symmetry as in the standard model. That again highlights how electroweak symmetry breaking and supersymmetry breaking may be connected. These masses also permit to shift all superpartners to such scales as they are in accordance with the observed limits so far. However, if the masses would be much larger than the scale of electroweak symmetry breaking, i. e., 250 GeV, it would again be very hard to obtain results in agreement with the observations without fine-tuning. However, such mass-matrices should not have large off-diagonal elements, i. e., the corresponding CKM-matrices should be almost diagonal. Otherwise, mixing would produce flavor mixing also for the standard model particles exceeding significantly the observed amount.

In addition, it is possible to introduce triple-scalar couplings. These can couple at will the squarks, sleptons, and the Higgs bosons in any gauge-invariant way. Again, observational limits restrict the magnitude of these couplings. Furthermore, some couplings, in particular certain inter-family couplings, are very unlikely to emerge in any kind of proposed supersymmetry breaking mechanism, and therefore are usually omitted.

Finally, the mass-matrices in the Higgs sector may have off-diagonal elements, without introducing mixing in the quark sector.

The arbitrariness of these enormous amount of coupling constants can be reduced, if some model of underlying supersymmetry breaking is assumed. One, rather often, invoked one is that the minimal supersymmetric standard model is the low-energy limit of a supergravity theory. In this case, it is, e. g., predicted, that the masses of the superpartners of the gauge-boson superpartners should be degenerate, and also the masses of the squarks and sleptons should be without mixing and degenerate and these masses also with the ones of the Higgs particles. Furthermore, all trilinear bosonic couplings would be degenerate. If the phase of all three parameters would be the same, also CP violation would not differ from the one of the standard model, an important constraint.

Of course, as the theory interacts, all of these parameters are only degenerate in such a way at the scale where supersymmetry breaks. Since the various flavors couple differently in the standard model, and thus in the minimal supersymmetric standard model, the parameters will again differ in general at the electroweak scale, or any lower scale than the supersymmetry breaking scale.

This simplest form of the MSSM is by now ruled out by experiment. To further permit a supersymmetric extension of the standard model requires to relax the assumptions. This can be either done by relaxing the supergravity constraints on the parameters, or by introducing additional new particles to rescue the supergravity paradigm. Either way introduces at the current time more unknowns, and further experimental data will decide whether this is he right way to go or whether supersymmetry is not part of nature, at least at energies not too large compared to the electroweak scale. However, generically the MSSM predicts a very light Higgs, and the observed mass is at the upper limit of the possible range. In fact, without quantum corrections due to rather heavy stops the Higgs mass is smaller than the Z mass. Thus, the outlook for the MSSM is somewhat tense.

Addition of further particles in so-called next-to-MSSM models alleviate these problems.

9.1.10 Supergravity

Because of the supersymmetry algebra (9.20), the generators are linked to the translation operator. In turn, gravity is the gauge theory of translations, as noted in section 7.1. Hence, making supersymmetry a gauge theory induces gravity, and vice versa. As the Nelson-Mandula theorem states that this is essentially the only possibility to do so, it is worthwhile exploring this a bit.

The metric, as a symmetric tensor, describes a spin 2 object. Supersymmetrizing gravity therefore requires a spin 3/2 field, which is the so-called Rarita-Schwinger field. This field will be the associated gauge field for the local supersymmetry, just as the metric field is the gauge field for the local translation symmetry.

In analogy to conventional gauge theories, the Rarita-Schwinger field is required to transform under a local transformation as

$$\Psi_\mu \to \Psi_\mu + \partial_\mu \epsilon$$

where ϵ is a spinor-valued function. Thus, a Rarita-Schwinger field Ψ carries both, a vector index and a spinor index. This was to be expected, as this couples effectively a spin 1 and a spin 1/2 object to create spin 3/2, just as for the metric two spin 1 indices are coupled to spin 2. Since the transformation is linear, it is an Abelian gauge theory, and the corresponding field strength tensor

$$\Omega_{\mu\nu} = \partial_\mu \Psi_\nu - \partial_\nu \Psi_\mu$$

is therefore gauge invariant, but carries also a spinor index.

It is still necessary to postulate a Lagrangian for the theory, which is gauge-invariant. Introducing $\bar{\Psi} = \Psi^\dagger \gamma_0$, a possibility is

$$\begin{aligned}
\mathcal{L} &= -\bar{\Psi}_\mu \gamma^{\mu\nu\rho} \partial_\nu \Psi_\rho . \\
\gamma^{\mu\nu\rho} &= \frac{1}{2} \{\gamma^\mu, \gamma^{\rho\sigma}\} \\
\gamma^{\mu\nu} &= \frac{1}{2} [\gamma^\mu, \gamma^\nu]
\end{aligned}$$

As for the Maxwell case, there are no gauge-invariant, perturbatively renormalizable further interaction terms possible. Without interactions, only non-interacting Rarita-Schwinger fields are possible. The equation of motion is, similar to the Dirac equation (3.4),

$$\gamma^{\mu\nu\rho} \partial_\nu \Psi_\rho = 0.$$

It follows that the Rarita-Schwinger field can have (classically) physical modes only for $d > 3$, similar like the vector potential only for $d > 2$. This equation of motion also implies

$$\gamma^\mu \Psi_{\mu\nu} = 0,$$

which is Rarita-Schwinger form of the Maxwell equations. The equations of motions can be solved in a similar way as the free Dirac equation, and creates the free-field solutions. It is possible to add a mass term, yielding

$$\mathcal{L} = -\bar{\Psi}_\mu(\gamma^{\mu\nu\rho}\partial_\nu - m\gamma^{\mu\rho})\Psi_\rho,$$

in contrast to the vector gauge fields.

The actual supergravity action is somewhat involved. Here, only the situation will be considered without additional matter fields, as they would have to be supersymmetrized as well. As shown above, the coupling of different matter multiplets leads to intertwining of those, which leads to a rather involved result. Also, the cosmological constant will be set to zero in the following.

The coupling between fermions and gravity actually is not a straightforward exercise in itself[7]. The approach taken here is based on exchanging the metric in favor of a different type of dynamical variables, the so-called vierbeins, defined as

$$g_{\mu\nu} = e_\mu^a \eta_{ab} e_\nu^b$$

where η is the space-time-constant Minkowski metric, and also the indices a and b run therefore from 0 to 3. This relation implies

$$e_a^\mu g_{\mu\nu} e_b^\nu = \eta_{ab},$$

i. e. the vierbein is the matrix field which yields a transformation of some given metric to the Minkowski metric. This field is therefore sometimes also called a frame field, as it locally transforms the metric to a Minkowski frame. Both indices of the vierbein can be raised and lowered using the Minkowski metric.

The Lagrangian of the simplest $\mathcal{N} = 1$ supergravity is then already quite complex,

$$\begin{aligned}
\mathcal{L} &= \frac{\det e}{2\kappa}\left(e^{a\mu}e^{b\nu}R_{\mu\nu ab} - \bar{\Psi}_\mu\gamma^{\mu\nu\rho}D_\nu\Psi_\rho\right)\\
R_{\mu\nu ab} &= \partial_\mu\omega_{\nu ab} - \partial_\nu\omega_{\mu ab} + \omega_{\mu ac}\omega_\nu{}^c{}_b - \omega_{\nu ac}\omega_\mu{}^c{}_b\\
D_\nu &= \partial_\nu + \frac{1}{4}\omega_{\nu ab}\gamma^{ab}\\
\omega_{\nu ab} &= 2e_{\mu[a}\partial_{[\nu}e_{b]}^{\mu]} - e_{\mu[a}e_{b]}^\sigma e_{\nu c}\partial^\mu e_\sigma^c
\end{aligned}$$

[7]In fact, there is more than one possibility, and they differ at the quantum level. Without experimental input, it is at the current time not possible to decide which one is correct. Here, the most prevalent construction is chosen.

where the brackets around the indices indicates that the expression has to be antisymmetrized with respect to the same-type indices. It is seen that the covariant derivative couples gravity and the Rarita-Schwinger field. This theory is therefore coupled. The involvement of the vierbein also makes the Einstein-Hilbert part more involved, and modifies the Riemann tensor, which now involves different types of indices.

The, now local, supersymmetry transformations of the fields are, without proof,

$$\begin{aligned}\delta e_\mu^a &= \frac{1}{2}\bar{\epsilon}\gamma^a\Psi_\mu\\ \delta\Psi_\mu &= D_\mu\epsilon,\end{aligned}$$

with the spinor 1/2 field $\epsilon(x)$.

Supergravity appears to allow for a perturbative quantization. However, especially its off-shell properties are much more involved. Still, what is known so far hints that it may be a viable theory. However, the implications of supersymmetry breaking as well as the absence of observation poiting towards supersymmetry remain a challenge.

9.2 Technicolor

Technicolor is the first prototype theory for compositness approaches. The idea is that the hierarchy problem associated with the mass of the Higgs boson can be circumvented if the Higgs boson is not elementary but a composite object. If its constituents in turn are made of particles with masses which do not suffer from a hierarchy problem, in particular fermions which have masses only affected logarithmically by perturbative quantum corrections, then the hierarchy problem simply would not exist.

However, such models require that interactions are non-perturbative, such that the Higgs can be a bound state. It would, as atoms, appear as an elementary particle only on energy scales significantly below the binding energy.

Such a construction is actually rather intuitive, and even realized in the standard model already. In QCD, bound states of quarks occur which actually have the same quantum numbers as the Higgs. In fact, already within QCD condensates with the quantum numbers of the Higgs condensate can be constructed, which induce the breaking of the electroweak symmetry. Only because the size of such condensates is then also given by the hadronic scale, and thus of order GeV, this is not sufficient to provide quantitatively for the observed electroweak symmetry breaking. Qualitatively it is the case.

Thus, the simplest extension is to postulate a second, QCD-like theory with typical energy scales in the TeV range with bound states which provide a composite Higgs, in addition to the standard model. Such theories are called technicolor theories.

Technicolor theories are also prototype theories for generic new strong interactions at higher energy scales, since at low energies they often differ from technicolor theories only by minor changes in the spectrum and concerning interaction strengths. Also, most of these suffer from similar problems as technicolor. Studying technicolor is therefore providing rather generic insight into standard model extensions with strongly interacting gauge interactions above the electroweak scale.

9.2.1 Simple technicolor

The simplest version of technicolor is indeed just an up-scaled version of QCD, though with a more general gauge group SU(N_T), with N_f additional fermions Q, the techniquarks. The techniquarks are massless at tree-level. They are placed in the fundamental representation of SU(N_T), and there are, in addition, the $N_T^2 - 1$ gauge bosons, called technigluons. Therefore, the total gauge group of the such extended standard model is SU(N_T)×SU(3)×SU(2)×U(1). The techniquarks harbor, similar to the ordinary quarks, a chiral symmetry. In such a theory the elementary particles include, besides the techniparticles, all the fermions and gauge bosons of the standard model, but no Higgs.

Such a theory then looks very much like QCD, though may have a different number of colors. Therefore, its dynamics are ought to be quite similar. In particular, technicolor confines, and techniquarks can only be observed bound in technihadrons. This dynamics will therefore be determined by a typical scale. In QCD, this scale is Λ_{QCD}, which is of order 1 GeV. This number is an independent parameter of the theory, and essentially replaces the coupling constant g_s of the elementary theory by dimensional transmutation. Now, in technicolor therefore there exists also such a scale Λ_T. To be of any practical use, it must be of the same size as the electroweak scale, otherwise the hierarchy problem will emerge again, though possibly less severe as a little hierarchy problem. Assume then that this scale is of the size 1 TeV instead of 1 GeV like in QCD. Then the dynamics of the technicolor theory would be the same as that of QCD, though at a much higher energy scale, and possibly with a different number of colors and flavors.

Besides the technimesons, which will play an important role in the electroweak sector as discussed below, there are also technibaryons. If they are fermionic, i. e. if N_T is odd, the lightest one can be stable, similar to the proton, and may thus exist as a remnant particle. However, in general these particles would be too strongly interacting, at least by quantum loop effects, than to be undetected by now. Hence, their decay into standard model particle is desirable, requiring a violation of the associated technibaryon number. If the number of technicolors N_T is even, the technibaryons are mesons, and could oscillate by mixing into mesonic states of the standard model, and would therefore decay definitely at least by such channels.

Another similarity with QCD is even more important. The techniquarks are so far massless. As in QCD, the chiral symmetry of the techniquarks is assumed to be broken spontaneously by the dynamics of the technigluons. The associated condensate will have a size of about Λ_T, and this gives the techniquarks an effective constituent mass of the order of Λ_T as well. Thus, technihadrons will have in general masses of multiple times the constituent techniquark mass. The only exception are the arising number of Goldstone bosons, similar to the pions and other pseudoscalar mesons of QCD. How many such Goldstone bosons appear depends on the number of techniflavors. In the present setup, their number will be N_f^2-1. As in QCD, these will be pseudoscalar bound states of a techniquark Q and an anti-techniquark $\bar{Q}$. If the techniquarks have the same weak and electromagnetic charges as the ordinary quarks, these technipions will just have the correct quantum numbers such that they can become the longitudinal components of the weak isospin W-bosons[8], instead of the would-be Goldstone bosons of the Higgs mechanism. Mixing with the hypercharge interaction will then lead as usual to the electroweak interactions. The Higgs is actually not one of the Goldstone bosons, but will be a scalar meson, the analogue of the σ meson of QCD. Thus, it is expected to be more massive, but also more unstable than the Goldstone mesons.

Note that for massless techniquarks the Goldstone bosons will be exactly massless. This can give rise to problems, as discussed below. However, it is not possible to give the techniquarks an explicit mass, because they have to be coupled chirally to the weak isospin. Thus, this remains a problem for $N_f>2$ in such simple technicolor theories, and how to resolve it will be discussed after illustrating other problems of this simplest setup.

The actual quantitative values for the various scales introduced can be estimated if the numbers of QCD are just scaled up naively to Λ_T, and the scaling to the number of technicolors N_T is done using the large-N_T approximation. The basic relation relates the electroweak condensate $v\approx 246$ GeV with the decay constant of the technipion. The later can then be related to the relation of the technicolor scale and the QCD scale with the pion decay constant in QCD f_{QCD}, which is measured to be about 93 MeV,

$$v=\langle\bar{Q}_L Q_R\rangle^{\frac{1}{3}}=f_T=\sqrt{\frac{N_T}{3}\frac{\Lambda_T}{\Lambda_{\text{QCD}}}}f_{\text{QCD}},$$

with the technichiral condensate $\langle\bar{Q}_L Q_R\rangle$. Solving for the technicolor scale yields

$$\Lambda_T\approx\sqrt{\frac{3}{N_T}}\frac{f_T}{f_{\text{QCD}}}\Lambda_{\text{QCD}}\approx\sqrt{\frac{3}{N_T}}0.7\text{ TeV}.$$

[8]Which already mix with the original pions, as pointed out before.

in a renormalization scheme (the $\bar{\mathrm{M}}\mathrm{S}$ scheme) with a Λ_{QCD} of about 250 MeV. Due to the breaking of the chiral symmetry, the effective mass of the techniquarks at lower energies is approximately also given by

$$m_Q(0) \approx v.$$

Though these are rather small masses, the techniquarks are not observable alone, similar to quarks at low energies. Thus, their direct detection is complicated by bound states, and their respective masses rather sets the scale for observation.

The mass of the Goldstone technipion is about

$$M_{\pi_T} \approx \sqrt{\frac{N_f}{2}}v,$$

and thus in the right region for them to be components of the W and Z bosons, if the number of flavors is not too large. Of course, QCD-like dynamics imply more bound states. Thus, masses of the low lying non-Goldstone bosons would start at about $2v \gtrsim 500$ GeV, plus binding effects. Assuming a QCD-like hierarchy, the next lightest state would be the techniρ, which would have a mass

$$M_{\rho_T} \approx \sqrt{\frac{3}{N_T}}\frac{vM_\rho}{f_{\mathrm{QCD}}} \approx \frac{3.3\text{ TeV}}{N_T^{\frac{1}{2}}},$$

and therefore would be sufficiently heavy to escape detection so far.

9.2.2 Extending technicolor

However, such a simple technicolor model has substantial problems. One of these additional problems is the following, indicating that the simplest technicolor models are not sufficient. This is that not only electroweak symmetry is broken by the Higgs, but also the fermion masses are generated by Yukawa couplings to the Higgs. To achieve this, the condensate of the techniquarks has to be coupled differently to all the standard model quarks. This is usually done by adding further massive gauge bosons with mass of a second scale $\Lambda_{\mathrm{ETC}} > \Lambda_T$, and coupled also to the standard model fermions. By this also the flavor group of the standard model becomes gauged in the extended technicolor gauge group, say, $\mathrm{SU}(\geq 6)_F\times\mathrm{SU}(N_T)\times\mathrm{SU}(3)\times\mathrm{SU}(2)\times\mathrm{U}(1)$. In general, this is achieved, by having a large extended technicolor (ETC) master gauge group, which contains all the other gauge groups, including the technicolor gauge group. All fermions, standard model ones and technifermions alike, are then embedded in representations of this master gauge group.

The breaking of the flavor group then provides a mechanism for the generation of the fermion mass, by their coupling to the now heavy flavor gauge bosons. However, despite giving a mechanism how the standard model fermion masses are generated, it is still a problem how to generate their relative sizes without the introduction of either new parameters or new fields. Note that this also explicitly breaks the chiral symmetry of QCD by fermion masses, as in the standard model.

Another problem is that the resulting effective couplings have to be very specific such that the hierarchy of fermion masses is obtained. E. g., quarks of mass a few GeV require a Λ_{ETC} of 2 TeV, while the top quark would rather require much less.

Nonetheless, such extended technicolor models are an important building block for promising technicolor theories, and therefore will be discussed here.

In general, the setting is to start with the master gauge group of extended technicolor G. It is broken by strong interactions at some scale Λ_{ETC} into the gauge group of one of the technicolor theories described previously. Then, at some lower scale Λ_T, the technicolor interactions become strong, leading to electroweak symmetry breaking as before. To avoid the hierarchy problem, it is often convenient not to make a single step from extended technicolor to technicolor, but have one or more intermediate steps, which in a natural way generate a hierarchy of scales. This is also known as a tumbling gauge theory scenario. The initial driving mechanism of the first breaking is not necessarily specified. A possibility would be that the master gauge group is part of a supersymmetric gauge theory, which provides naturally a hierarchy-protected Higgs mechanism, as an initial starting point.

A variation on this theme are triggering models. In this case the fact that QCD breaks the electroweak symmetry is used to plant a seed of breaking also an extended technicolor gauge group. This seed is then amplified by a suitable arrangement of interactions such that the right hierarchy of scales emerge. This can also be done with other triggers than QCD.

The extended technicolor gauge bosons, which have become massive on the order of Λ_{ETC} still interact with all particles. In particular, they mediate interactions purely between techniparticles, between ordinary standard model particles and techniparticles, and only between standard model particles. Similarly to the electroweak interactions, they are perceived at the scale Λ_T and below as effective four-fermion couplings. Since the technifermions condense, the mixed couplings have a contribution which couple the standard-model particles to the condensate, schematically

$$g^2\bar{T}\gamma_\mu T D^{\mu\nu}_{\mathrm{ETC}}(p^2)\bar{q}\gamma_\nu q \overset{p^2\ll\Lambda_{\mathrm{ETC}}}{\to} \frac{g^2}{M_{\mathrm{ETC}}^2}\bar{T}\gamma_\mu T\bar{q}\gamma^\mu q \overset{p^2\ll\Lambda_T}{\to} \frac{g^2\langle\bar{T}T\rangle}{M_{\mathrm{ETC}}^2}\bar{q}q,$$

with extended technicolor coupling constant g, the mass of the extended technigluon M_{ETC}, and the techni fermion condensate $\langle\bar{T}T\rangle$. Thus, the techniquark condensate indeed also generates the masses of the standard model fermions on top of the W and Z boson masses.

The size of the quark masses is given approximately by

$$m_q \approx \beta \frac{N_T \Lambda_T^3}{\Lambda_{\mathrm{ETC}}^2} \tag{9.25}$$

where the factor β depends on the structure of the theory. For $\beta \approx 1$, a 'natural' size, this yields an upper limit on Λ_{ETC} from the masses of the light quarks, to be about not much more than an order of magnitude larger than Λ_T, and this only if Λ_T is not too large itself. On the other hand, if Λ_{ETC} should not be too large, this is an upper bound for the quark masses, which can be produced. In fact, if Λ_T is not much smaller than one to two orders of magnitude than Λ_{ETC}, and N_T is not too large, even achieving the bottom quark mass, much less the top quark mass, is hardly possible.

An advantage of such an interaction is that, depending on the detailed structure of the interactions, this can even provide some of the mixing of the standard model CKM and PMNS matrices. However, the same holds for the effective four-techniquark coupling, coupling states of two techniquarks also to the techniquark condensate. This gives rise to larger masses for the technigoldstone bosons. Unfortunately, the size of this effect is essentially given by the ratio of the extended technicolor and the technicolor scale. If both scales are not very far apart, the effect is unfortunately not large enough to give those particles too light in technicolor theories a sufficiently large mass to be compatible with experimental bounds. Increasing the extended technicolor scale is not a solution, since this spoils the masses of the light standard model fermions. A solution to this will be the walking technicolor theories discussed in the next section.

A further downside is that also a coupling between four standard model fermions is induced, which contribute to, e. g., flavor-changing neutral currents. As a result, e. g., the mass difference between the two kaon states δm_K^2, called short and long, is modified by ETC contributions to

$$\frac{\delta m_K^2}{m_K^2} \to \frac{\delta m_K^2}{m_K^2} + \gamma \frac{f_K^2 m_K^2}{\Lambda_{\mathrm{ETC}}^2}.$$

Herein γ is the effective coupling between standard model quarks. If assumed to be of the same size as the Cabibbo angle, which mediates the mixing in the standard model and is of order 10^{-2}, Λ_{ETC} has to be of order 10^3 TeV, which clearly exceeds the expected size. It is one of the persisting challenges of extended technicolor theories to provide at the same time the mass of the top quark without having such currents to be so large that they are in conflict with experiments. Actually, this problem also affects other beyond-the-standard-model theories, most notably supersymmetry. It thus makes evident that one of the greatest challenges is to understand the flavor structure of the standard model.

A further problem is that such an interaction yields corrections to quantities like the coupling of the Z boson: If the Z boson is first converted into a techniquark pair, and then these convert via extended technigluon exchange into ordinary standard model gauge bosons, this will yield vertex corrections which are essentially given by ratios of the technicolor scale, giving the coupling to the Z boson of the techniquarks, and the extended technicolor scale, relevant for the conversion ratio of techniquarks to standard model fermions. Since this ratio is not too large, the corrections are significant, and indeed ruled out.

Thus, a challenge is to construct extended-technicolor models, which are nonetheless consistent with observations. Some candidates will be discussed below. Note that in general the introduction of elementary scalars, which can have varying technicolor and standard model charges, can remove most of the problems. The advantage compared to the Higgs boson of the standard model, for which to remove technicolor was invented in the first place, is that these Higgses can have a rather large mass, as they only contribute partly to the electroweak effects, the rest coming from technicolor. Therefore, they can be embedded in a higher-scale theory, like a supersymmetric one, where their masses become protected by additional symmetries from a hierarchy problem. This appears to be a valid alternative in case supersymmetric particles are not found, but strong interactions in weak gauge boson scattering indicate a strongly interacting theory as the origin of electroweak symmetry breaking. In particular, even if these scalars do not condense, they can mediate additional interactions between the technicondensates, removing several of the large effects incompatible with the experimental observations in technicolor models, including the top mass.

9.2.3 Conformal field theories

To resolve the contradiction between the large top quark mass and flavor physics will require to introduce an additional effect, which is called walking. The idea of walking is based on the properties of conformal field theories. It is therefore quite useful to briefly introduce conformal field theories, and discuss their most pertinent features.

Conformal theories have a space-time symmetry which is much larger than the ordinary Poincare symmetry. In particular, they are also symmetric under two additional transformations,

$$\begin{aligned} x^\mu &\to \lambda x^\mu \\ x^\mu &\to \frac{x^\mu + a^\mu x^2}{1 + 2x^\nu a_\nu + a^2 x^2}, \end{aligned}$$

with arbitrary vector a_μ. The first is a scale transformation, which is also called dilation. The second is the special conformal transformation. Together with the Poincare algebra these two transformations form the conformal algebra. The scale symmetry is in so far remarkable, as only scaleless theories can be conformal. Hence, a conformal theory never has any kind of intrinsic mass scale.

Conformal symmetry is hence a very strong constraint, which restricts the correlation functions of a theory severely. E. g., the two-point correlator of any two fields in a conformal theory is bound to behave as

$$\langle\phi(0)\phi(x)\rangle \sim \frac{1}{(x^2)^\Delta},$$

with the scaling exponent Δ fixed by the scaling dimension of the field ϕ under a conformal transformation $x \to \lambda x$

$$\phi(x) \to \lambda^\Delta \phi(\lambda x).$$

Similarly, the three-point functions are restricted to behave as

$$\langle\phi_1(x_1)\phi_2(x_2)\phi_3(x_3)\rangle \sim \frac{c_{ijk}}{|x_1-x_2|^{\Delta_1+\Delta_2-\Delta_3}|x_1-x_3|^{\Delta_1+\Delta_3-\Delta_2}|x_2-x_3|^{\Delta_2+\Delta_3-\Delta_1}}.$$

In two dimensions, these constraints are so strong that this permits to solve any theory exactly, as the exponents are uniquely fixed, and there are descent equations for the coefficients. The reason is that the conformal symmetry introduces an infinite number of conserved quantities in two dimensions. In higher dimensions, the constraints are less severe, and solving conformal theories is in general not possible anymore. However, the symmetry still restricts the behavior of the theory rather rigidly.

The most important constraint of the conformal symmetry in context of particle physics already stems from the scale symmetry. Since this symmetry forbids the appearance of a mass scale, the running coupling (3.11) cannot have a scale Λ. Hence, the only possibility is that the running coupling (and all other running quantities) do not run, but are constant as a function of energy. This implies that the β-function (3.10) vanishes identical.

Hence, in a conformal theory, there is no difference between large and small energy. Thus, such theories do not have any particles in the ordinary sense, as the do not have a mass. As a consequence, any cross-section must be energy-independent, and thus there are no measurements in the ordinary sense. A conformal theory has no physics, and thus cannot describe nature. But it is mathematically interesting, and will serve as an important guideline for the construction of physically relevant technicolor theories. Especially, adding to a conformal theory a small, but explicit, breaking of conformality can lead to quite interesting models. Many technicolor models probably belong to this class, but proving exact conformality in the underlying theory is not always entirely trivial.

As a side-remark, the probably most famous conformal theory in four dimensions is an extension of the massless supersymmetric Yang-Mills theory introduced in section 9.1.5.

By adding further massless scalars and fermions and restricting the couplings in a very particular way such that only the gauge coupling remains independent, it is possible to add three more independent supersymmetries to the theory. These supersymmetries act in different ways on the different particles. They are not entirely different, as the algebra of all four symmetries are connected to the momentum operator. In fact, the four supersymmetries can be transformed into each other by an additional, ordinary, symmetry, the so-called R-symmetry[9]. Hence, this theory is known as $\mathcal{N} = 4$ Super-Yang-Mills theory. Leaving these technicalities aside, in this theory the contributions from the different particles and interactions cancel just so as to make the β function of the gauge coupling vanish, and thus the theory is conformal. In fact, this is called superconformal.

9.2.4 Walking technicolor

The basic reason why a similarity of technicolor to QCD is problematic is that in QCD almost all non-trivial dynamics is concentrated in a narrow window around Λ_{QCD}. That is because the running coupling of QCD changes rather quickly from strong to weak over a very narrow range of energies. In the electroweak sector, however, the dynamics is spread out over a much larger range of energy scales in relation to its fundamental scale v, since both the masses of the fermions and electroweak symmetry breaking must occur. Thus, a viable realization of the technicolor idea of strong dynamics paired with electroweak phenomenology must reflect the slow evolution of the electroweak physics. This is the aim of walking technicolor by replacing the fast running QCD evolution with a much slower, walking, behavior.

As a consequence, such a theory has more intrinsic scales than QCD. QCD is essentially only characterized by the one scale when it becomes strong. A walking theory can have up to three scales. Assuming the walking theory to be also asymptotically free, there exists a scale where it changes from being a theory acting strongly enough to break electroweak symmetry to an almost free theory. A second scale must occur at low energy when it stops walking, and the third scale is the one where it becomes sufficiently strong to confine techniquarks. Of course, the latter two may coincide, but the first two may not, or the theory would no longer be walking anymore.

For a guideline how to construct such theories conformal theories now come into play. Conformal theories are the paradigmatic example of energy-independent theories. As a downside, they lack dynamics. The compromise of walking theories is then to have a theory which, for some range as indicated above, is almost conformal. This suggests how

[9] R parity is a discrete version of it.

to construct such theories: Such theories must have β-functions, which are close to zero, but not exactly zero.

That this is indeed giving the desired behavior can be seen from a rewriting of (3.10),

$$t=\int\limits_{g}^{g(t)}\frac{dg'}{\beta(g')},$$

where $g=g(\Lambda_T)$ is some chosen initial conditions, and $t=\ln\mu/\Lambda_T$. Hence, if the β function is approximately zero, the momentum dependency is very slow.

The general β function to third order of perturbation theory in a QCD-like setup is given by

$$\begin{aligned}
\beta(g) &= -\beta_0\frac{g^3}{16\pi^2}-\beta_1\frac{g^5}{(16\pi^2)^2}-\beta_2\frac{g^7}{(16\pi^2)^3}+\mathcal{O}(g^9)\\
\beta_0 &= \frac{11}{3}C_A-\frac{4}{3}N_fT_R\\
\beta_1 &= \frac{34}{3}C_A^2-\frac{20}{3}C_AT_RN_f-4C_RT_RN_f\\
\beta_2 &= \frac{2857}{54}C_A^3-\frac{1415}{27}C_A^2N_fT_R+\frac{158}{27}C_A(N_fT_R)^2\\
&\quad -\frac{205}{9}C_AC_FN_fT_R+\frac{44}{9}C_F(N_fT_R)^2+2(C_F)^2N_fT_R.
\end{aligned}$$

Here, N_f is the number of flavors in the fundamental representation, i. e. the same as the quarks, and T_R is the so-called Dynkin index of the group. It is depending on the conventions, but in the context of particle physics it is usually 1/2. C_A is a second Casimir invariant of the group, and equals the number of colors N for an SU(N) group.

If the function β is very close to zero for some value of g, $g(t)$ becomes a very slowly varying function of t when it reaches this value. E. g., to leading order of the β function of technicolor with techniquarks in the fundamental representation of an SU(N_T) techni gauge group with gauge coupling g_T this yields

$$\beta_T=-\frac{g_T^3}{16\pi^2}\left(\frac{11}{3}N_T-\frac{8}{3}N_f\right)+\mathcal{O}(g^5)$$

requiring about $N_f\approx 11N_T/8$. Note that the standard model charges of these techniquarks are not relevant at this order. Therefore, by judiciously choosing the gauge group and the number of flavors, it is possible to construct a β function giving a theory that has the desired walking behavior. More possibilities are offered by exchanging the representation of the techniquarks. Using instead of fundamental techniquarks adjoint techniquarks, i. e.

quarks with the same color charges as the gluons, shows that for $N_T = 2$ the β function with $N_f = 2$ already vanishes to two-loop order[10].

It should be noted that this argumentation can only be superficial: The β functions is dependent on the renormalization scheme, and the running coupling can be defined in many ways. It is therefore a much more subtle task to indeed show that a theory is walking than the outline discussed here. But it points to the correct direction that it just requires a careful balancing of the various degrees of freedom in the theory.

The big advantage of such a theory is that the coupling evolves slowly with energies. Therefore, the theory stays strong over a wide range of energies. As a consequence, techni bound states can no longer spoil various electroweak precision measurements. Furthermore, when arranging this walking behavior for the range between Λ_T and Λ_{ETC}, the interaction of the standard model fermions among each other and mediated by the extended technigluons will essentially be independent of energy, and thus remain small, while the electroweak dynamics is only given by the technicolor dynamics and stays essentially unaltered. In fact, what happens is that (9.25) is modified to

$$m_q \approx \gamma \frac{N_T \Lambda_T^2}{\Lambda_{\text{ETC}}} \tag{9.26}$$

and thus quark masses of order one to two GeV are possible for reasonably chosen values of Λ_T and Λ_{ETC} of 1 and a few tens of TeV.

A similar replacement also takes place for the masses of the Goldstone boson masses which are not absorbed by the W and Z bosons. Their mass is now found to be of order $N_T \Lambda_T$, and thus sufficiently large to be not detectable yet.

A variation on the idea of walking technicolor is given by low-scale technicolor. In this case the techniquarks necessary to make the theory walk are in different representations of the gauge group. Since thus their respective energy scales are different, the corresponding condensates, which add up quadratically to form the electroweak condensate, form at different energies. These scales are widely separated due to the walking behavior. As a consequence, techni bound states could have masses of the same size as the top quark, though being sufficiently weakly coupled to escape detection so far.

However, even with the relation (9.26) this is only marginally sufficient to obtain the bottom quark mass, and for the top quark an excessively fine-tuned value of γ would be required.

---

[10] In fact, most recent investigations point to the possibility that this is even too effective, making g not only slowly varying with t but even independent of t, and the theory may even become conformal.

9.3 Grand unified theories

9.3.1 Setup

As outlined before, the fact that the electromagnetic couplings have ratios of small integers for quarks and leptons cannot be explained within the standard model. However, this is necessary to exclude anomalies, and make the standard model a valid quantum field theory. Thus, this oddity suggests that possibly quarks and leptons are not that different as the standard model suggests. The basic idea of grand unified theories is that this is indeed so, and that at sufficiently high energies an underlying symmetry relates quarks and leptons, making these ratios of electric charges natural. This is only possible, if the gauge interactions, and thus the gauge group $\mathrm{SU}(3)_{\mathrm{color}}\times\mathrm{SU}(2)_{\mathrm{weak}}\times\mathrm{U}(1)_{\mathrm{em}}$ is also embedded into a single group, since otherwise this would distinguish quarks from leptons due to their different non-electromagnetic charges. Another motivation besides the electromagnetic couplings for this to be the case is that the running couplings, the effective energy dependence of the effective gauge couplings, of all three interactions almost meet at a single energy scale, of about 10^{15} GeV. They do not quite, but if the symmetry between quarks and leptons is broken at this scale, it would look in such a way from the low-energy perspective. If all gauge interactions would become one, this would indeed require that all the couplings would exactly match at some energy scale.

These arguments are the basic idea behind grand unified theories (GUT) to be discussed now for a simple (and already by experiment excluded) example. Since there are very many viable options for such grand-unified theories, all of which can be made compatible with what is known so far, there is no point as to give preference of one over the other, but instead just to discuss the common traits for the simplest example. Also, GUT ideas are recurring in other beyond-the-standard model scenarios. E. g., in supersymmetric or technicolor extensions the required new parameters are often assumed to be not a real additional effect, but, at some sufficiently high scale, all of these will emerge together with the standard model parameters from such a GUT. In these cases the breaking of the GUT just produces further sectors, which decouple at higher energies from the standard model. Here, the possibility of further sectors to be included in the GUT will not be considered further.

The basic idea is that such a GUT is as simple as possible. The simplest version compatible with just the structure of the standard model requires to have a Yang-Mills theory with a single, simple gauge group, and the matter fields belong to given representations of it. As noted, the standard model gauge group is $\mathrm{SU}(3)_{\mathrm{color}}\times\mathrm{SU}(2)_{\mathrm{weak}}\times\mathrm{U}(1)_{\mathrm{em}}$. This is a rank 4 Lie group, 2 for SU(3) and 1 for SU(2) and U(1). Thus, at least a group of rank 4 is necessary, excluding, e. g., SU(4) with rank 3 or G_2 with rank 2. Furthermore, fermions are described by complex-valued spinors, and thus complex-valued representations must exist. This would be another reason against, e. g., G_2, which has only real representations. Another requirement is that no anomalies appear in its quantization.

Taking everything together, the simplest Lie groups admissible are SU(5) or SO(10), both having rank 4, as well as the rather popular cases of rank 6, 7, and 8, the groups E_6, E_7, and E_8, respectively.

Now, take SU(5) for example. It has 24 generators, and thus 24 gauge bosons are associated with it. Since the standard model only offers 12 gauge bosons, there are 12 too many. These can be removed when they gain mass from a Brout-Englert-Higgs effect, if the masses are sufficiently large, say of order of 10^{15} GeV as well. Thus, in addition to new heavy gauge bosons, a number of additional Higgs fields, or other mediators of symmetry breaking, are necessary in GUTs.

It should be noted that the idea of GUTs in this simple version, i. e., just be enlarging the gauge group, cannot include gravity. It requires, e. g., the addition of supersymmetry or another structural change to allow this, as disucssed in section 9.1.

9.3.2 A specific example

Lets take as a specific example SU(5) for the construction of a GUT. It has 24 generators, and therefore there are 24 gauge bosons. 8 will be the conventional gluons G_μ, 3 the W_μ of the weak isospin bosons, and 1 the hypercharge gauge boson B_μ, leaving 12 further gauge bosons. These 12 additional gauge bosons can be split in four groups of three X, Y, $X^\dagger$, and $Y^\dagger$, making them complex in contrast to the other gauge bosons for convenience. The structure requires then that the gauge bosons X_μ and Y_μ carry electric charge 5/3, while their complex conjugate partners have the corresponding anti-charge $-5/3$. In much the same way it can be shown that the three elements of each of the four fields can be arranged such that these gauge bosons carry the same color and weak isospin as the (left-handed) quarks and leptons. Since they can therefore couple leptons and quarks directly, they are referred to as leptoquarks, mediating e. g. proton decay as discussed below.

Arranging the fermions turns out to be a bit more complicated. Each family consists of 16 fermionic particles, 12 quarks and 4 leptons, counting two quark and lepton flavors with three colors for the quarks and left-handed and right-handed chiralities separately. The fundamental representation of SU(5) is only of dimension 5, and can therefore not accommodate this number of particles. Also, the assignment in multiple copies of the fundamental representation cannot yield the correct quantum numbers. Therefore, the matter fields must be arranged in a non-trivial way.

It is an exercise in group theory, not to be repeated here in detail, that the simplest possibility is to assign the 16 particles to three different multiplets. In this construction the right-handed neutrino ν_R become a singlet under SU(5). Since already in the standard model it only couples to the remaining physics with a strength measured by its very small mass, and therefore not at all at high energies, this appears appropriate. The remaining particles of a family are put in two further multiplet structures. The right-handed down quarks and left-handed electron and electron neutrino can be put into an anti-5 (anti-fundamental) multiplet ψ while the remaining particles can be arranged in a 10-multiplet χ. This multiplet structures also ensures an anomaly-free theory.

This appears to be a quite awkward way of distributing particles, and also not be symmetric at all. However, without proof, this distribution yields that all fermions have the correct quantum numbers. In particular, the correct electric charges are assigned - and exactly those. Hence, this embedding implies the mysterious relation of quark and lepton charges of the standard model in a natural way. Furthermore, it also implies that right-handed quarks are not interacting weakly, in this sense yielding parity violation of the weak interactions as well. It is thus a so-called left-right symmetric theory, and parity violation becomes just a low-energy effect due to the breaking of the gauge group.

The remaining problem is now the presence of the X and Y gauge bosons. At the current level, these are massless. Even if they would be coupled to the Higgs field of the standard model, this would have to occur in the same way as with the W and Z bosons, thus yielding approximately the same masses. That is in contradiction to experiments, and therefore some way has to be found to provide them with a sufficiently heavy mass as to be compatible with experiments.

The simplest possibility is to have a second Brout-Englert-Higgs mechanism, like in the electroweak sector. Since the latter may not be affected, two sets of Higgs fields are necessary. The simplest possibility is to have one multiplet of Higgs fields $\Sigma = \Sigma^a \tau^a$ in the 24-dimensional adjoint representation of SU(5), and another one H in the fundamental five-dimensional one. The prior will be used to break the SU(5) to the unbroken standard model gauge group, and the second to further break it to the broken standard model. This provides an additional mass shift for the Y bosons, and for the W and Z bosons their usual standard model masses. Out of the 10 independent degrees of freedom in the fundamental representation 3 are absorbed as longitudinal degrees of freedom of the W and Z bosons, leaving seven Higgs bosons. One of them has the quantum numbers of the standard model Higgs boson, while the other six decompose into two triplets (like quarks) under the strong interactions. By introducing appropriate couplings between Σ and H bosons, it is possible to provide these six with a mass of the same order as the leptoquarks, thus also making them inaccessible at current energies.

It remains to show how fermion masses are protected from becoming also of order w. Actually, it is not possible to construct a coupling between Σ and the fermions which is renormalizable. However, it is possible to construct a Yukawa coupling to the H Higgs bosons. At the GUT scale, these require the same mass for down quarks and electrons, and similar relations for the other quarks and leptons. Transferring these results to the scale of the Z mass yields results which are in good agreement with experiment for some mass ratios, notably the bottom-to-τ ratio is about three, close to the experimental value of 2.4. However, in particular the light quark masses are not obtained reasonably well, showing that this most simple GUT is not sufficient to reproduce the standard model alone.

9.3.3 Running coupling

After this very specific example, it is also possible to make some more general statements, which will be done in this and the next section.

One of the motivation to introduce a grand-unified theory was the almost-meeting of the running couplings of the standard model when naively extrapolated to high energies. Because of the requirement that the structure of GUTs should be simple, all matter fields couple with the same covariant derivative to the GUT gauge bosons. Here, it will be assumed that each generation of standard model matter fields fill exactly one (or more) multiplet(s) of the theory, but no further additional particles are needed to fill up the multiplets, and no multiplets contain particles from more than one generation.

In one-loop approximation a generic coupling constant g changes with energy μ, see (3.11), as

$$\left(\frac{g(\mu)^2}{4\pi}\right)^{-1} = \alpha(\mu)^{-1} = \alpha(\Lambda_{\mathrm{GUT}})^{-1} - \frac{\beta_0}{2\pi}\ln\frac{\Lambda_{\mathrm{GUT}}}{\mu}, \tag{9.27}$$

where Λ_{GUT} is the energy scale were the gauge symmetry of the GUT is manifest. The only missing ingredient are then the β functions, which are specific for the respective coupling, and can be calculated, e. g., in perturbation theory. For a generic manifest gauge theory, like QCD, coupled to fermions and scalars it is given by

$$\beta_0 = \frac{11}{3}C_A - \frac{2}{3}T_f - \frac{1}{3}T_s,$$

where C_A is the adjoint Casimir of the gauge group, e. g. 5 for SU(5), and T_f and T_s depend on the number and representation of fermions and scalars, respectively, with masses below μ. If below the GUT scale only the known particles (including a light Higgs) exist, then the relevant coefficients β_0 are given by

$$\begin{aligned}
\beta_0^s &= \quad 11 - \tfrac{4}{3}N_g &&= 7\\
\beta_0^i &= \tfrac{22}{3} - \tfrac{4}{3}N_g - \tfrac{1}{6}N_H &&= \frac{19}{6}\\
\beta_0^h &= 0 - \tfrac{20}{9}N_g - \tfrac{1}{6}N_H &&= -\frac{41}{6},
\end{aligned}$$

for the three interactions, where $N_g = 3$ is the number of fermion generations and $N_H = 1$ the number of complex Higgs doublets. As an aside, note that a unification of the couplings would be impossible if the electric interaction would be stronger than either or both of the strong and weak gauge coupling, the later only appearing small by the effective masses of the W and Z bosons, since it is the only running coupling which increases with energy. Based on preference, this can also be taken as an indication for unification.

It then follows that the Weinberg θ_W angle behaves as

$$\sin^2\theta_W = \frac{3}{8} - \frac{\alpha_i^{-1} + \alpha_h^{-1}}{2\pi}\frac{109}{24}\ln\frac{\Lambda_{\text{GUT}}}{\mu}.$$

To eliminate the unknown scale Λ_{GUT} another evolution equation can be used. Particularly convenient is the combination

$$\alpha_i^{-1} + \alpha_h^{-1} - \frac{8}{3}\alpha_s = \frac{67}{6\pi}\ln\frac{\Lambda_{\text{GUT}}}{\mu},$$

yielding

$$\sin^2\theta_W = \frac{23}{134} + \frac{\alpha_i^{-1} + \alpha_h^{-1}}{\alpha_s}\frac{109}{201}.$$

Using the experimental values $\alpha_i^{-1} + \alpha_h^{-1} = 128$ and $\alpha_s = 0.12$ at the Z-boson mass, $\mu = M_Z$ yields $\Lambda_{\text{GUT}} \approx 8 \times 10^{14}$ GeV, $\alpha_s(\Lambda_{\text{GUT}}) \approx 1/42$, and $\sin^2\theta_W(M_Z) = 0.207$. The latter number is uncomfortably different from the measured value of 0.2312(2), implying that at least at one-loop order this GUT proposal is not acceptable.

Unfortunately, this problem is not alleviated by higher-order corrections, and turns out to be quite independent of the particular unification group employed, and many other details of the GUT. This implies that unification cannot occur with the simple setup discussed here. Only when other particles, in addition to the minimum number needed to realize the GUT, are brought into play with masses between the electroweak and the GUT scale a perfect unification can be obtained. Supersymmetry, e. g., provides such a unification rather naturally. Of course, the fact that Λ_{GUT} is much closer to M_P than to the electroweak scale could also be taken as a suggestion that quantum gravity effects may become relevant in the unification process. These are open questions.

9.3.4 Baryon number violation

A reason not to easily abandon GUT theories after the disappointment concerning the running couplings is that they naturally provide baryon number violation, which is so necessary to explain the matter-antimatter asymmetry of the universe. That such a process is present in GUTs follows immediately from the fact that quarks and leptons couple both to the same gauge group as one multiplet. Thus, gauge bosons can mediate transformations between them, just as the weak gauge bosons can change quark or lepton flavor individually. Whether some quantum numbers are still conserved depends on the details of the GUT. A GUT with gauge group SU(5), e. g., preserves still the difference of baryon number B and lepton number L, $B-L$. For the gauge group SO(10), not even this is conserved.

The profound consequence of baryon and lepton number violation is the decay of protons to leptons. It is in principle a very well defined experimental problem to measure this decay rate, though natural background radiation makes it extremely difficult in practice. To estimate its strength, assume for a moment that the masses of the gauge bosons mediating this decay are much heavier than the proton, which is in light of the experimental situation rather justified. Then the decay can be approximated by a four-fermion coupling, very much like a weak decay can be approximated by such a coupling at energies much smaller than the masses of the mediating W and Z bosons.

The life-time at tree-level can then be calculated in standard perturbation theory in leading order to be

$$\tau_{p\to e^+\pi^0}\sim\frac{192\pi^3}{G_{\mathrm{GUT}}^2 m_p^5}\sim\frac{m_{\mathrm{GUT}}^4(\alpha_i^{-1}(m_{\mathrm{GUT}})+\alpha_h^{-1}(m_{\mathrm{GUT}}))}{m_p^5}.$$

Plugging in the previous numbers, the formidable result is about 10^{31} years. That appears quite large, but the current experimental limit for this channel is about 10^{34} years, clearly exceeding this value. Thus, at least this very simple approximation would yield that a GUT is in violation of the experimental observation by about three orders of magnitude. However, pre-factors and higher order corrections depend very much on the GUT under study, and can raise the decay time again above the experimental limit. Finding proton decay, or increasing further the limit, would provide therefore information on the structure of permitted GUTs. However, the search becomes experimentally more and more challenging, so that pushing the boundaries further is an expensive and demanding challenge. Nonetheless, this is a worthwhile problem: If no proton decay should be observed ere reaching the standard model decay rate, this would imply that cosmological baryon number violation would proceed in a rather unexpected way.

9.4 Other low-scale extensions of the standard model

In the following briefly some other possibilities to add particles to the standard models are presented, which still adhere to a four-dimensional space time and ordinary quantum field theories without gravity. In contrast to the theories discussed later, these require less drastic changes at the electroweak or 1 TeV scale to our current picture of nature.

9.4.1 Little Higgs

The idea of the Higgs as an emergent state is also the primary guideline for the construction of little Higgs models. If the Higgs would be the Goldstone boson of a broken global symmetry, it would naturally be light, in fact even massless if the symmetry-breaking would be only spontaneous, similar to the pion in QCD. The simplest case would be an additional global symmetry with some particles charged under it, which becomes broken at the TeV scale.

However, such a simple model is usually inappropriate, and more refined approaches are necessary. One of them is the idea of collective symmetry breaking. To become more formal, such physics is usually described using a non-linear sigma model

$$\mathcal{L} = \frac{1}{2}\partial_\mu\phi\partial^\mu\phi + \frac{1}{2}\frac{(\phi\partial_\mu\phi)(\phi\partial^\mu\phi)}{|f^2-\phi^2|}. \tag{9.28}$$

If f is zero, this reduces to a free scalar theory. This Lagrangian can be linearized to the linear sigma model by the introduction of another field σ to

$$\mathcal{L} = \frac{1}{2}\partial_\mu\Phi\partial^\mu\Phi - \frac{\lambda}{4}\left(\Phi\Phi - \frac{\mu^2}{\lambda}\right)^2. \tag{9.29}$$

where $\Phi = (\phi, \sigma)$. If the symmetry is broken, $f^2 = \phi^2 + \sigma^2$ and f is a function of the parameters λ and μ. The field ϕ is massless, and plays the role of the Goldstone boson, here the Higgs boson. The original Lagrangian (9.29) is invariant under a symmetry group G acting on Φ, while the effective Lagrangian is only so under a smaller symmetry group G/H acting on ϕ. Thus, to specify the non-linear sigma model, a strength f and a symmetry breaking pattern $G \to G/H$ is necessary. At an order $4\pi f \sim \Lambda$, the effective description in terms of (9.29) breaks down, as then the energy is sufficiently large to excite σs.

To achieve decent agreement with experiment is challenging with this concept. It is necessary to take a group G, with a gauged subgroup SU(2)$\times$U(1) to obtain a Higgs with correct properties. Also, there must be explicit breaking, which can be modeled by a mass-term $m^2\phi^2$ in (9.28). Though this approach gives a first possibility, it turns out that it endows a (little) hierarchy problem, since the emerging Higgs is again having a mass sensitive to corrections at the scale Λ.

As a remedy, the mentioned collective symmetry breaking was introduced. The basic idea is to use a product group $G_1 \times G_2$ which has a gauged electroweak group SU(2)×U(1) in each of the factor groups G_i. In such a setup, radiative mass corrections between both factor groups actually cancel, at least at (one-)loop level, such that the Higgs mass is protected. However, to obtain reasonable masses for the top quark and the Higgs simultaneously requires an additional vectorial, i. e. coupling to weak interactions without parity violation, partner of the top quark, usually denoted by T. Then the top quark can have a large mass, without its (large) Yukawa coupling to the Higgs leading to large radiative corrections of its mass, since the latter are canceled by contributions from its T partner.

Such subtle cancellations are a hallmark of the various little Higgs models. Popular examples for the choice of G are the minimal moose model $(\mathrm{SU}(3)_L \times \mathrm{SU}(3)_R/\mathrm{SU}(3)_V)^4$, where even a SU(3)×SU(2)×U(1) subgroup is gauged, leading to additional gauge bosons which become heavy by the symmetry breaking, the littlest Higgs model with SU(5)/SO(5) having a gauged $(\mathrm{SU}(2)\times\mathrm{U}(1))^2$ subgroup, and the simplest little Higgs model with group G being $(\mathrm{SU}(3)\times\mathrm{U}(1)/\mathrm{SU}(2))^2$ with gauged subgroup SU(3)×U(1).

However, in all cases some fine-tuning appears at some point to obtain results in agreement with experimental data. A radical approach to remedy the problem is by introducing an additional global Z_2 symmetry, under which standard model and additional particles are differently charged. The action of this symmetry is to exchange the subgroups G_i of G. This is called T parity. Provided T-parity is not developing an anomaly, what indeed happens for some models, the lightest additional particle is stable, and thus a dark matter candidate.

9.4.2 Hidden sectors

Of course, not every proposal for new physics is necessarily designed to solve one or more of the problems of the standard model. It would be a naive expectation that only such new physics could emerge. Hence, a wide variety of scenarios have been investigated for additional physical effects, especially with focus on whether they can be added to particle physics without upsetting the available experimental data.

A rather generic and characteristic example are hidden sectors. The idea of hidden sectors is that in addition to the standard model there is a second set of particles which have very weak or no coupling to the standard model particles, and a set of very heavy messenger particles connecting this hidden sector to the standard model. Provided that these particles are not gravitationally bound in significant numbers to ordinary astrophysical objects, such a sector will not be detectable unless the energies reached become of order of the messenger masses.

A simple example for a hidden sector would be a hidden QCD with some gauge group[11] SU(N) and hidden quarks charged under this symmetry. The mediator is a U(1), i. e., QED-like symmetry with a gauge boson Z'. This symmetry is broken at the TeV scale, making the Z' very heavy. If the hidden quarks also have mass of this size, but the hidden QCD is unbroken, a high-energetic Z' can be produced by standard model particles, and then decay into a hidden hadron, which decays to the lightest state, generically a hidden pion, which can then decay through a virtual Z' to standard model particles.

Another possibility is the quirk scenario. In this case the hidden quarks, called quirks, are in addition also charged like standard model quarks, but very heavy compared to the intrinsic scale Λ_{hidden} of the hidden gauge theory. Then the quirks themselves act as mediators. As a consequence, hidden glueballs would be quasi-stable on collider time-scales, giving unique missing energy signatures.

Another possibility is if the hidden theory is almost conformal, and only coupled weakly to the standard model. In this case the conformal behavior of the hidden particles will generate very distinctive signatures, as their kinematic behavior is quite distinct from anything the standard model offers on these energy scales. Such hidden (quasi-)conformal particles are also called unparticles.

9.4.3 Flavons

A serious obstacle in technicolor theories had been the generation of the mass spectrum of the fermions. To remedy this problem, a hidden sector can be introduced.

In this case all Yukawa couplings of the standard model are dropped, i. e., all fermions are exactly massless. Then there exists an additional global symmetry, the flavor symmetry, which has symmetry group U(3)^6. Some part of it is broken by QCD due to chiral symmetry breaking, generating most of the mass for the up, down and strange quark, but (almost) nothing for leptons, and not enough for the heavier quarks. To provide it, quarks and leptons are coupled to a further field, called flavon, by a photon-like messenger particle. Neither are charged under the standard model gauge interactions. The flavon then condenses, and the messenger couples the condensate back to the standard model fermions, providing their masses. Though this is not explaining the mass hierarchy, this splits the dynamics of the fermion mass generation from electroweak symmetry breaking,

[11] N has to be larger than two for compatibility with experiment.

which could then, e. g., be provided by a rather simple technicolor theory. The messenger becomes unobservable, since the flavon condensates provides it with a heavy mass. The messenger needs to be a U(1)-type gauge boson, to permit different couplings to the different flavors.

Integrating out the messenger field will generate couplings between the Higgs (or whatever replaces it) and the standard model fermions, which will essentially look like the standard-model Yukawa couplings. The Yukawa couplings will then be proportional to the ratio of the electroweak condensate squared and the mass of the messenger particle. Assuming the lightest particle, neutrinos, to have a coupling to the messengers of order 1 then yields a mass for the messenger of order Λ_{GUT}, and therefore further consequences of these particles will not be harmful to present electroweak precision measurements. This scenario is also known as the Froggat-Nielsen mechanism.

9.4.4 See-saw

One of the most puzzling properties of the mass hierarchy in the standard model is the enormously small masses of the neutrinos. To explain this different mass scale the so-called see-saw mechanism can be used.

This mechanism operates by introducing one or more additional neutrinos χ_i, which mix with the neutrinos. Consider the case of a single neutrino flavor and one additional partner. A gauge-invariant mass matrix for the neutrino and its χ partner can then be written as

$$(\bar{\nu}_L\bar{\chi}_L)\begin{pmatrix}0 & g_\nu v\\ M_{\nu\chi} & M_{\chi\chi}\end{pmatrix}\begin{pmatrix}\nu_R\\ \chi_R\end{pmatrix}+\mathrm{h.c.},$$

where $M_{\nu\chi}$ and $M_{\chi\chi}$ are free parameters, v is the Higgs condensate, and g_ν the Yukawa coupling to the Higgs, which now may be of order one. Mass eigenstates can be obtained by diagonalization. The obtained masses for the actual mass eigenstates are then the eigenvalues given by

$$m_i=\frac{1}{2}\left(M_{\nu\chi}^2+M_{\chi\chi}^2+(g_\nu v)^2\pm\sqrt{\left(M_{\nu\chi}^2+M_{\chi\chi}^2+(g_\nu v)^2\right)^2-4g_\nu^2v^2M_{\nu\chi}^2}\right) \tag{9.30}$$

Chosen appropriately, the lighter of the two eigenstates acquires the mass of the standard-model neutrino, while the other is much more heavier, and can have easily a mass in or above the TeV range. Expanding the masses in this case gives

$$\begin{aligned}m_1 &= \frac{g_\nu^2v^2M_{\nu\chi}^2}{M_{\nu\chi}^2+M_{\chi\chi}^2+(g_\nu v)^2}+\mathcal{O}\left(\frac{g_\nu^4v^4M_{\nu\chi}^4}{M_{\chi\chi}^6}\right)\\ m_2 &= M_{\nu\chi}^2+M_{\chi\chi}^2+(g_\nu v)^2+\mathcal{O}\left(\frac{g_\nu^2v^2M_{\nu\chi}^2}{M_{\chi\chi}^4}\right)\end{aligned}$$

For sufficiently large $M_{\chi\chi}$ and not too large $M_{\nu\chi}$, the one state is much lighter, and experimentally suitable masses can be obtained. In this way, the observed very light neutrinos are no longer different from the the other fermions of the standard model.

9.4.5 Axions and other light particles

There is a certain possibility that even very light particles are existing, and remain undetected, if these particles couple sufficiently weakly to the standard model. This can be achieved by being either only coupled through a hidden sector messenger, or by only coupling electromagnetically and exploiting that a U(1) structure permits arbitrary small charges for every particle independently. Such particles are often referred to as axions.

Axions have originally been invented to provide a dynamical explanation why certain additional terms do not appear in the standard model Lagrangian though being permitted by all other constraints. Most notable of these terms is one which would introduce a CP violation, in addition to the flavor mixing, through QCD, and a similar one in the weak interaction. Experimentally, this additional contribution has been constrained to be less than 10^{-10} that of the CP conserving part of the strong force, and in the weak sector it should already be from the outset almost unmeasurablely small. Why it is so small is unclear. Axions usually implement a Higgs-like effect, which makes the contribution dynamically zero or close to zero. At the same time, they are only coupled through this effect, or, as noted, by other weak effects. Thus, they can be very light. Since they can be bosons, they can nonetheless be very common, and therefore are dark matter candidates. Furthermore, they can contribute to problems where additional energy transport is required.

The are several variations of the axion theme, some of them also being detached from the original strong CP problem. Generically, they therefore refer to weakly coupled, light particles, and are actively searched for. Of course, since they are weakly coupled, they could be also heavy. There are various acronyms for the different flavors of the axion idea, like WIMP for weakly-interacting massive particles, WISP for weakly-interacting scalar particles, and many more.

9.4.6 2-Higgs doublet models

There is a generic trait for many BSM scenarios: The appearance of additional scalar particles, being them elementary or composite. All of these models have a very similar low-energy behavior, essentially the standard model with more Higgs-like particles. This whole class of models is hence known as 2-Higgs doublet models. And of course, these include any models which have this structure from the outset. The different high-energy theories manifest then themselves in the various parameters of this extended Higgs sectors, e. g., whether one or both Higgs doublets condense, and what kind of additional symmetries are present. Thus, such theories are heavily studied.

9.5 At the Planck scale

9.5.1 Large extra dimensions

One attempt to remedy the large difference in scale between gravity and the standard model is by the presence of additional dimensions. Also, string theories, to be discussed later, typical require more than just four dimensions to be well-defined. Such extra dimensions should not be detectable. The simplest possibility to make them undetectable with current methods is by making them compact, i. e., of finite extent. Upper limits for the extensions of such extra dimensions depend strongly on the number of them, but for the simplest models with two extra dimensions sizes of the order of micrometer are still admissible. These large extra dimensions are contrasted to the usually small extensions encountered in string theory, which could be of the order of the Planck length. For now, start with those with large extension. These can also be considered as low-energy effective theories of, e. g., string theories.

These models separate in various different types. One criterion to distinguish them is how the additional dimensions are made finite, i. e. how they are compactified. There are simple possibilities, like just making them periodic, corresponding to a toroidal compactification, or much more complicated ones like warped extra dimensions. The second criterion is whether only gravity can move freely in the additional dimensions, while the standard model fields are restricted to the uncompactified four-dimensional submanifold, then often referred to as the boundary or a brane in contrast to the full space-time called the bulk, or if all fields can propagate freely in all dimensions.

One thing about these large extra dimensions is that they can also be checked by tests of gravity instead of collider experiments. If there are $4+n$ dimensions, the gravitational force is given by

$$F(r) \sim \frac{G_N^{4+n} m_1 m_2}{r^{n+2}} = \frac{1}{M_s^{n+2}} \frac{m_1 m_2}{r^{n+2}},$$

where G_N^{4+n} is the $4+n$-dimensional Newton constant and correspondingly M_s the $4+n$-dimensional Planck mass. If the additional n dimensions are finite with a typical size L, then at large distances the perceived force is

$$F(r) \sim \frac{1}{M_s^{n+2}L^n}\frac{m_1m_2}{r^2} = \frac{G_N m_1 m_2}{r^2},$$

with the four-dimensional Newton constant G_N. Thus, at sufficiently long distances the effects of extra dimensions is to lower the effective gravitational interactions by a factor of order L^n. That can explain why gravitation is perceived as such a weak force at sufficiently long distances.

On the other hand, by measuring the gravitational law at small distances, deviations from the $1/r^2$-form could be detected, if the distance is smaller or comparable to L. This is experimentally quite difficult, and such tests of gravity have so far only been possible down to the scale of the order of a little less than a hundred μm. If the scale M_{4+n} should be of order TeV, this excludes a single and two extra dimensions, but three are easily possible. Indeed, string theories suggest n to be six or seven, thus there are plenty of possibilities. In fact, in this case the string scale becomes the $4+n$-dimensional Planck scale, and is here therefore denoted by M_s. The following will discuss consequences for particle physics of these extra dimensions.

9.5.1.1 Separable gravity-exclusive extra dimensions

The simplest example of large extra dimensions is given by theories which have n additional space-like dimensions, i. e., the metric signature is $\mathrm{diag}(-1, 1, ..., 1)$. Furthermore, these additional dimensions are taken to be separable so that the metric separates into a product

$$g^{4+n} = g^4 \times g^n.$$

Furthermore, for the additional dimensions to be gravity exclusive the other fields have to be restricted to the 4-dimensional brane of uncompactified dimensions. In terms of the Einstein equation (7.5) this implies that the total energy momentum tensor T_{MN} takes the form

$$T_{MN} = \begin{pmatrix} T_{\mu\nu} & 0 \\ 0 & 0 \end{pmatrix}, \tag{9.31}$$

where the indices M and N count all dimensions and μ and ν only the conventional four. Furthermore, in such models the extra dimensions are compact, having some fixed boundary conditions.

The Einstein-Hilbert action is then

$$S_{EH} = -\frac{1}{2G_{N4+n}^{1+n/2}}\int d^{4+n}z\sqrt{|g^{4+n}|}R^{4+n} \tag{9.32}$$

with again the generalized Newton constant G_{N4+n}, the metric g and the Ricci scalar R. The action then factorizes as

$$S_{EH} = -\frac{M_s^{n+2}}{2}\int d^{4+n}z\sqrt{|g^{4+n}|}R^{4+n} = -\frac{1}{2}M_P^2\int d^4x\sqrt{-g^4}R^4.$$

The actual gravity mass-scale M_s is then related to the perceived 4-dimensional Planck scale by

$$M_P = M_s(2\pi RM_s)^{\frac{n}{2}} = M_s\sqrt{V_nM_s^n},$$

with the volume of the additional (compact) dimensions V_n, which have all the same compactification radius R. For an M_s of order 1 TeV, the compactification radius for $n=2$ to $n=6$ ranges from 10^{-3} m to 10^{-11} m, being at $n=2$ just outside the experimentally permitted range.

Treating the theory perturbatively permits to expand the metric as

$$g_{MN} = \eta_{MN} + \frac{2}{M_s^{1+\frac{n}{2}}}H_{MN},$$

with the usual Minkowski metric $\eta_{AB} = \mathrm{diag}(-1,1,...,1)$ and the metric fluctuation field H_{AB}. The Einstein-Hilbert action is then given by an integral over the Einstein-Hilbert Lagrangian

$$\mathcal{L}_{EH} = -\frac{1}{2}H_{MN}\partial^2H^{MN}+\frac{1}{2}H_N^N\partial^2H_M^M-H^{MN}\partial_M\partial_NH_L^L+H^{MN}\partial_M\partial_LH_N^L-\frac{1}{M_s^{1+\frac{n}{2}}}H^{MN}T_{MN}.$$

Since the additional dimensions are finite, it is possible to expand h_{MN} in the additional coordinates in a series of suitable functions f_n, embodying the structure of the extra dimensions

$$\begin{aligned} H_{MN}(x_0,..,x_3,x_4,...,x_{3+n}) &= \sum_{m_1,...,m_n} f_n(k^4_{m_1,...,m_n}x_4+...k^{3+n}_{m_1,...,m_n}x_{3+n})H_{MN}(x_0,...,x_3) \\ k_{m_1,..,m_n} &= \left(K\left(\frac{\pi m_1}{R}\right),...,K\left(\frac{\pi m_n}{R}\right)\right)^T \end{aligned}$$

The energies k can be related in the usual way to masses, $k_m^2 = m_{KKm}^2$, as in quantum mechanics. These are called the Kaluza-Klein masses. The field h_{MN} for a fixed mass can then be decomposed into four four-dimensional fields. These are a spin-2 graviton field $G_{\mu\nu}$, $i=1,..,n-1$ spin-1 fields[12] A_μ^i, $i=1,...,(n^2-n-2)/2$ scalars S^i and a single further scalar h. Soft modes are the zero-modes of the Fourier-transformed fields, i. e., those with $m_{KK}^2=0$. The fields A and S do not couple to the standard model via the

[12]Originally, Kaluza and Klein in the 1930s aimed at associating this field with the electromagnetic one, which failed.

energy momentum tensor, and the graviton coupling is suppressed by the Planck mass, in agreement with the observation that gravity couples weakly. This also applies for the radion h, which couples to the trace of the energy-momentum tensor, corresponding to volume fluctuations. However, because of its quantum numbers, it will (weakly) mix with the Higgs.

Finally, since $m_{KKn} \sim k_n \sim n/R$ for a mode n, the level splitting of the Kaluza-Klein modes is associated to the size of the extra dimensions. The splitting is thus given by

$$\delta m_{KK} = m_{KKn} - m_{KKn-1} \sim \frac{1}{R} \approx 2\pi M_s \left(\frac{M_s}{M_P}\right)^{\frac{2}{n}}.$$

which is generically of order meV for $n = 2$ to MeV for $n = 6$. Thus, to contemporary experiments with their limited resolution of states the Kaluza-Klein states will appear as a continuum of states.

A serious problem arises when the universally coupling Kaluza-Klein modes show up in processes forbidden, or strongly suppressed, in the standard model, like proton decay. The standard model limit for proton decay by an effective four fermion vertex is about 10^{15} GeV, thus much larger than the comparable effect from the larger extra dimensions if M_s should be of order TeV. Thus this leads to a contradiction if not either M_S (or n) is again set very large, and thus large extra dimensions become once more undetectable, or additional custodial physics is added to this simple setup. The latter usually leads, like in the case of technicolor, to rather complex setups.

9.5.1.2 Universal extra dimensions

The alternative to gravity-exclusive extra dimensions are such which are accessible to all fields equally. This implies that the theory is fully Poincare-invariant prior to compactification in contrast to the previous case. As a consequence, such theories can in general not resolve the hierarchy problem. However, they provide possibilities how anomalies can be canceled, e. g. in six dimensions, without need to assign specific charges to particles. In addition, one of the Kaluza-Klein modes can often serve as a dark matter candidate. On the other hand, since particle physics has been tested to quite some high energy with no deviations observed, this imposes severe restrictions on the size of extra dimensions, being usually of order inverse TeV, and thus sub-fermi range, rather than μm.

When compactifying the additional dimensions in such theories care has to be taken when imposing the boundary conditions. The reason is that fermions in a box with anti-periodic boundary conditions will develop an effective mass of order $1/L$, where L is the compactification scale. This is due to the fact that only odd frequencies in a Fourier-expansion of a fermion field have the right periodicity, $(2n + 1)L$, instead of $2nL$ as for bosons, as required by the spin 1/2 nature of fermions. Therefore, chiral boundary conditions are required. These are conditions which permit zero frequency even for odd oscillations, e. g. open boundary conditions. The mass-spectrum of all standard model particles for a compactification along a single extra dimension with open (chiral) boundary conditions, a so-called orbifold, is then given by

$$M_j^2 = \frac{\pi^2 j^2}{L^2} + m_0^2,$$

where m_0 is the mass of the standard model particle, and its Kaluza-Klein excitations have mass M_j, and j counts the excitation.

9.5.1.3 Warped extra dimensions

In models with warped extra dimensions, also known as Randall-Sundrum models, the additional dimensions have an intrinsic curvature k in such a way that the energy scales depend exponentially on the separation Δy of two points in the additional dimensions, $\exp(-2\Delta yk)$. By positioning different effects at different relative positions, large scale differences can appear naturally, e. g., $M_H \sim \exp(-\Delta yk)M_P$. In particular, the different Yukawa couplings for the standard model fermions can be explained by having different wave functions for different fermion species in the additional dimension, which then have different overlap with the Higgs wave function, therefore permitting very different couplings to the Higgs, even if the difference is of order unity in a flat space. This is very similar to the concept of split fermions in the case of universal extra dimensions.

In the minimal version of warped extra dimensions there is only one additional dimension. This one is orbifolded, i. e., it is compactified on a radius πR with opposite points identified[13], giving the additional coordinate y the range from 0 to πR. The fifth dimension is bounded by two end-points, which are four-dimensional. These two end-points are called branes. Due to the explicit exponential, both are not the same, but differ by a metric factor of $\exp(2k\pi R)$. One of the branes is identified with our four-dimensional world. Its Planck mass is then related to the Planck mass of the bulk M_g, i. e., inside the total volume, by

$$M_P^2 = \frac{M_g^3}{k}\left(1 - e^{-2\pi kR}\right).$$

A natural size for k is about M_g itself, and it is also natural to have $kR \gtrsim 1$. Then the Planck mass and the bulk Planck mass are again of the same magnitude, despite that otherwise only natural scales appear.

[13] Topologically, this is S^1/Z_2.

Why there is nonetheless no discrepancy between the electroweak and the Planck scale is explained thus differently in such models than in the large extra dimensional models beforehand. Take the brane at $y = \pi R$ to be our world. Assume that the Higgs H is confined to this brane. It is then described by the action

$$\begin{aligned} S &= \int d^4x\sqrt{|g(y=\pi R)|}(g_{\mu\nu}(y=\pi R)(D^\mu H)^\dagger(D^\nu H) - \lambda(HH^\dagger - V^2)^2) \\ &= \int d^4x(\eta_{\mu\nu}(D^\mu H)^\dagger(D^\nu H) - \lambda(HH^\dagger - e^{-2\pi kR}V^2)^2), \end{aligned}$$

where in the second step the Higgs field has been rescaled by $H \to \exp(\pi kR)H$, to remove the exponential from the four-dimensional induced metric. As a consequence, the expectation value of the Higgs is $\langle H\rangle = \exp(-\pi kR)V = v$, and by this a quantity naturally of the same scale as the Planck mass is scaled down to the much smaller electroweak scale by the exponential pre-factor. To have the correct numbers for a V of the size of the Planck scale $kR \approx 11$ is needed. This solves the hierarchy problem, or actually makes it nonexistent.

So far, all the standard model fields have been restricted to the infrared brane. Permitting further particles to also propagate in the fifth dimension requires some subtle changes. Gauge fields will then have five components, instead of four. Choosing an appropriate gauge and Dirichlet boundary conditions[14] make the fifth component vanish altogether. Physically, the absence of the fifth component of the field can then be interpreted as that this component provides the necessary longitudinal degree of freedom for the massive Kaluza-Klein gauge bosons. Unfortunately, as in the graviton case, the Kaluza-Klein modes couple enhanced by a factor $k\exp(\pi kR)$ to the standard model particles as in case of the graviton. This can only be avoided at the cost of having the geometry such that all new physics is moved to rather large energies, reintroducing the hierarchy problem, or by rather subtle manipulations on the kinetic terms of the gauge bosons on the ultraviolet brane.

This changes, if also the fermions can propagate into the bulk. However, this is again complicated by the chiral nature of fermions. Chirality of five-dimensional fermions is fundamentally different from the one of four-dimensional ones. This can be remedied by introducing a second set of fermions with opposite chirality of the standard model ones. To avoid that all of them are visible, it is necessary to give them different boundary conditions. Only fields with von Neumann boundary conditions[15] on both branes are found to have (up to the Higgs effect) massless modes, and can therefore represent the standard

[14] I. e. fixed values for the additional fields on the branes.

[15] I. e. specifying the derivatives of the additional fields on the branes.

model fermions. Fermions with Dirichlet boundary conditions on at least one boundary immediately acquire a Kaluza-Klein mass. Therefore, they will not be visible below the TeV scale. Such scenarios require quite a number of amendments, like extra symmetries or particles, to make them compatible not only qualitatively but also quantitatively with experimental precision measurements.

9.5.1.4 Symmetry breaking by orbifolding

With the orbifolds it is also possible to provide symmetry breaking. Take for example the SU(5) GUT of section 9.3.2. In this case it was necessary to introduce numerous additional Higgs fields to remove the additional gauge bosons, acting as additional leptoquarks, from the spectrum to have a decent proton life time. This can also be achieved by orbifolded extra dimensions. Take for example a single extra dimension with boundary conditions.

Take the SU(5) gauge field. It is splitted into the standard model fields of gluons, W, Z, and the photon. Furthermore the X and Y gauge fields appear. In addition, universal extra dimensions dictate to have a further fifth component for all gauge fields. The fifth dimension is different than the ordinary four dimensions by its different structure. In addition to ordinary parity transformations in four dimensions, there is also an exclusive five-dimensional parity transformation $y \to -y$. The fifth component of the gauge field must then have opposite parity under $y \to -y$. This can be seen from the fact that ∂_y is necessarily odd under $y \to -y$. The component $F_{\mu 5} = \partial_\mu A_5 - \partial_5 A_\mu$ of the field-strength tensor must have a definite parity under this transformation, or otherwise the theory would not respect the orbifold structure of the theory. Thus, A_μ and A_5 have opposite boundary conditions.

Choosing then Neumann boundary conditions for the standard model fields automatically makes their fifth component having Dirichlet boundary conditions, making them heavy. To also remove the X and Y gauge bosons together with their fifth component from the low-energy realm requires them to have mixed boundary conditions. By this, the GUT symmetry appears to be explicitly broken since all additional fields have become massive. This is the concept of symmetry breaking by orbifolding.

9.5.1.5 Deconstructed extra dimensions

An alternative flavor of (large) extra dimensions are obtained from so-called deconstructed extra dimensions. In this case the extra dimensions are not continuous, but are discrete, i. e., contain only a finite number of points, like a lattice. This removes the ultraviolet divergences encountered by having an infinite number of Kaluza-Klein states, making the theory renormalizable. This can also be viewed by a finite, in case of the extra dimension being compactified, or infinite set of four-dimensional space-times, which are distinguished by a discrete quantum number.

As an example, take only one additional dimension, with N points and radius R. Then each of the N points is a complete four-dimensional space-time, and is also called a brane. Take now a gauge-field $A_\mu(x, y)$ with y the (discrete) fifth coordinate. On a given brane, the fifth coordinate is fixed and denotes the brane. There are then four gauge-field components depending on the remaining four coordinates x, just as a normal gauge field would. This can, e. g., give the gluons of QCD. There is another field, the fifth component of the gauge field, depending on a fixed brane again on the four coordinates. It can be shown to behave like an adjoint Higgs field.

Expanding the gauge field in a discrete Fourier series shows the presence of further, heavier Kaluza-Klein modes as copies of these fields. From a low-energy perspective, like an experiment, these appear in addition to the gauge theory described by the zero modes as $N-1$ copies of this gauge theory, which are broken by the additional $N-1$ adjoint Higgs fields, giving the Kaluza-Klein modes of the gauge fields their mass. The remaining zero-mode of the A_5 component can be rearranged such that it can take the role of the standard model Higgs, breaking the electroweak symmetry[16].

Similarly, it is possible to introduce fermions having the correct chiral properties by choosing appropriate boundary conditions, as before. As a bonus, tuning the parameters appropriately, it is possible to make Kaluza-Klein fermions condense, essentially realizing a so-called topcolor mechanism, i. e. a replacement of the Higgs condensate and the Higgs itself by top-anti-top states, and thus providing the mechanism usually unspecified in topcolor theories.

9.5.1.6 Black holes

A rather popular possible signature for large extra dimensions are the production and decay of black holes. The Schwarzschild radius of a $4+n$-dimensional black hole for n compact dimensions characterized by the $4+n$-dimensional Planck scale M_s is given by

$$R_B \sim \frac{1}{M_s}\left(\frac{M_B}{M_s}\right)^{\frac{1}{n+1}},$$

with the black hole mass M_B. If in a high-energy collisions two particles with center-of-mass energy s larger than M_s^2 come closer than R_B, a black hole of mass $M_B \approx s$ is

[16]Theories exploiting the same mechanism to obtain the standard model Higgs for a continuous extra dimensions are sometimes called gauge-Higgs unifying theories or also holographic Higgs theories.

formed. The cross-section is thus essentially the geometric one,

$$\sigma \approx \pi R_B^2 \sim \frac{1}{M_s^2}\left(\frac{M_B}{M_s}\right)^{\frac{2}{n+1}}.$$

It therefore drops sharply with the scale M_s. However, its decay signature is quite unique. It decays by the Hawking process, i. e., by the absorption of virtual anti-particles by the black hole, making their virtual partner particles real. The expectation value for the number of particles for the decay of such a black hole is

$$\langle N\rangle \sim \left(\frac{M_B}{M_s}\right)^{\frac{n+2}{n+1}},$$

and therefore rises quickly when the energies of the colliding particles, and thus the mass of the produced black hole, significantly exceeds the scale of the compactified dimensions.

9.5.2 Quantum gravity

In the following a short look will be taken at how a quantum theory of gravity can be obtained pursuing the normal way of covariantly quantizing it. There are other possibilities, like using a path integral as in section 8.2.2, as well. All of this will be rather formal compared to the rest of the lecture. The reason is that a consistent theory of quantum gravity is very hard to obtain. Because the Newton constant appearing is of dimension of inverse energy any perturbative way of quantizing the theory will produce a non-renormalizable theory, having little predictivity. Therefore, ultimately, a genuine non-perturbative approach will be necessary. Thus, here more the formal aspects will be discussed, which, however, will in the end yield that the graviton has to be massless.

Quantum gravity as a direct quantized version of general relativity is also not the only proposed possibility of how to obtain classical gravity in the long-distance limit of everyday life and of stellar or interstellar objects. Other approaches include local supersymmetry leading to supergravity theories, string theories, or non-commutative geometry. All of them have their merits and drawbacks, but all are also measured in comparison to quantum gravity. A first encounter should therefore give here a flavor of how quantum gravity looks and what problems are encountered with it. The basic setup of string theory in the next section can then be compared to it.

9.5.2.1 Quantization

The invariance under general coordinate transformation makes the (canonical) quantization of gravity complicated and akin to the quantization of a gauge theory, but with the coordinate transformations taking the place of the gauge transformation. A practical implementation is therefore even more complicated than for a gauge theory, and will be skipped here. It should be noted, however, that there is no unique way yet known of how to quantize gravity, i. e. on which quantities e. g. canonical commutation relations should be imposed. This is actually not surprising. Even for ordinary quantum mechanics, there is no unique choice. It was rather the comparison to experiment which eventually decided what was the correct way to proceed. Quantization rules are postulates, not derived.

Given that there are currently no experimentally reliable statements about quantum effects in gravity, this option is not yet available. Hence, the only possibility is to quantize the theory, and obtain its low-energy limit, which must be classical general relativity, for which plenty of experimental and observational results exist. This problem becomes even more severe once fermions are included, as these are true quantum objects without classical analogue.

As a consequence, several different proposals are currently pursued. One is the canonical quantization approach, which essentially imposes canonical commutation rules on the local fields, i. e. the metric tensor. When doing so, it turns out that the quantized theory is only diffeomorphism-invariant, if the metric tensor, which is now the field of the graviton, is massless. The consequence of this quantization procedure is that also the coordinates loose their meaning in the usual way. Since their values are subject to quantum fluctuations of the metric, it is no longer possible to assign to them an independent reality, as has been done in the classical theory. Thus the quantum uncertainty not only afflicts the particles but space-time as such: Not only it becomes impossible to assign a position to a particle, but it becomes impossible to give the term position a meaning. Thus, even if particles could be assigned a position, the position concept itself has now a quantum uncertainty.

Though the quantization procedure is in itself, irrespective of the particular version chosen, straightforward, the resulting quantum theory is not.

If quantized canonically, the major obstacle is that the theory cannot be treated perturbatively, as noted already in section 8.2.1. A possible solution is that this is only a technical problem, and the theory is treatable once non-perturbative methods are employed. This is the concept of asymptotic safety, discussed in section 8.2.2.

9.5.2.2 Loop quantum gravity

In other versions of quantization this is not necessarily true. One particular popular choice is the so-called loop quantum gravity approach. In this case, the quantization is not applied to the elementary field, the metric, itself, but rather to diffeomorphism-invariant quantities, which are certain closed line integrals of the metric, therefore the reference to loops. It appears that this quantization procedure leads to a perturbatively more tractable theory, but the inherent non-localities when defining the dynamical variables have so far limited the progress.

9.5.2.3 Non-commutative geometry

One further possibility to quantize gravity is to postulate the existence of a minimal length, similar to the postulate of a minimal phase space volume $\Delta x \Delta p \sim \hbar$ in ordinary quantum mechanics. This is also similar to the idea of a maximum speed in general relativity. As there, the existence of such a minimal length, which is typically of the order of the Planck length 10^{-20} fm, has profound consequences for the structure of space-time. Especially, coordinate operators do no longer commute, just like coordinate and momenta do not commute in quantum mechanics. Thus, the same effect can be reached by postulating canonical commutation relations for coordinates, in addition to the ones between coordinates and momenta. Thus, this ansatz is called non-commutative geometry. Since there is a minimal length, there is also a maximal energy, and hence all quantities become inherently finite, and renormalization is no longer necessary. On the downside of this approach, besides an enormous increase in technical complexity, is that in general relativity neither coordinates nor energies themselves are any longer physical entities, like in special relativity or in quantum (field) theories. Thus, the precise physical interpretation of a non-commutative geometry is not entirely clear. Furthermore, so far it was not possible to establish a non-commutative theory which, in a satisfactory manner, provides a low-energy limit sufficiently similar to the standard model. Particularly cumbersome is that it is very hard to separate the ultraviolet regime where the non-commutativity becomes manifest and the infrared, where the coordinates should again effectively commute.

9.5.3 String theory

A complete alternative to the quantization of conventional gravity is string theory. Here, the problems encountered when attempting to quantize the Einstein-Hilbert Lagrangian are transcended by using a conceptual new form of theory, which only in the low-energy limit takes the form of an ordinary quantum field theory. In these theories the concept of an elementary particle, i. e. one-dimensional objects, are replaced by strings, i. e. two-dimensional objects.

However, most string theories have a very complicated structure, and it is particularly non-trivial to understand the low-energy limit. This problem is made more severe by the fact that a given string theory can have multiple low-energy limits, which are hard to distinguish, the so-called landscape problem. Furthermore, string theories are technically complicated, especially when treated consistently. Hence, here only a brief glimpse can be given of their properties. Especially, any string theory which seems to be able to produce nature requires at some scale a (broken) supersymmetry, leading to superstring theories. Finally, all string theories found so far are related to each other. Furthermore, it turns out that they all need common objects, which are of higher-dimensional types than strings, the already alluded to branes. The description of branes and strings is conjectured to be unifiable in a single theory, the so-called matrix theory, or M-theory for short. Until now, it was not possible to formulate this theory.

Fortunately, string theories have a number of rather generic properties, like the natural appearance of gravitons, the need for additional dimensions, and problems due to tachyons, which appear already in their simplest form. It is therefore possible to illustrate these features already in the quantum mechanics version of string theory. This will be done in this section.

This already shows, why string theories are particularly attractive: The natural appearance of the graviton makes string theories rather interesting, given the intrinsic problems of quantum gravity. Further advantages of more sophisticated string theories are that they have generically few parameters, are not featuring space-time singularities, such as black holes, on a quantum level, and often have no need for renormalization, thus being consistent ultraviolet completions. The price to be paid is that only rather complicated string theories even have a chance to resemble the standard model, their quantization beyond perturbation theory is not yet fully solved, and it is unclear how to identify a string theory which has a vacuum state which is compatible with standard model physics, again the so-called landscape problem due to enormous number of (classical) vacuum states in string theories. Furthermore, in general genuine string signatures usually only appear at energy levels comparable to the Planck scale, making an experimental investigation, and thus support, of stringy properties of physics, almost impossible with the current technology.

How comes the idea of string theory about? It is motivated by how perturbation theory fails. Generically, the understanding of physics has been increased by looking at ever shorter distances and at ever high energies. The current confirmed state of affairs is then the standard model. Going back to quantum gravity, a similar insight can be gained. In the perturbative approach, the ratio of free propagation to a tree-level exchange of a graviton is essentially given by the interaction strength of gravity times a free graviton propagator, which is essentially given by the inverse of $G_N E^2$, with E the energy of the graviton. Thus the corresponding ratio is

$$\frac{A_{\text{free}}}{A_{1g}} = \frac{\hbar c^5}{G_N E^2} = \frac{M_P^2}{E^2}$$

where $M_P^2 = \hbar c^5 / G_N$ is again the Planck mass, this time in standard units. Since M_P is once more of the order of 10^{19} GeV, this effect is negligible for typical collider energies of TeV. However, if the energy becomes much larger than the scale, the ratio of free propagation to exchange of a graviton becomes much smaller than one, indicating the breakdown of perturbation theory.

This is not cured by higher order effects. E. g., in case of the two-graviton exchange, the corresponding amplitude ratio becomes

$$\frac{A_{2g}}{A_{\text{free}}} \sim (\hbar G_N)^2 \sum_{\text{Intermediate states}} \int_0^E dE' E'^3 \sim \frac{1}{M_P^4} \int dE' E'^3 \to \infty \text{ for } E \to \infty \tag{9.33}$$

This gets even worse with each higher order of perturbation theory. Thus, perturbation theory completely fails for quantum gravity. Either non-perturbative effects kick in, or something entirely different. That might be string theory.

The basic idea behind string theory is to try something new. The problem leading to the divergence of (9.33) is that with ever increasing energy ever shorter distances are probed, and by this ever more gravitons are found. This occupation with gravitons is then what ultimately leads to the problem. The ansatz of string theory is then to prevent such an effect. This is achieved by smearing out the interaction over a space-time volume. For a conventional quantum field theory such an inherent non-locality usually comes with the loss of causality. String theories, however, are a possibility to preserve causality and smear out the interaction in such a way that the problem is not occurring.

However, the approach of string theory actually goes (and, as a matter of fact, has to go) a step further. Instead of smearing only the interaction, it smears out the particles themselves. Of course, this occurs already anyway in quantum physics by the uncertainty principle. But in quantum field theory it is still possible to speak in the classical limit of a world-line of a particle. In string theory, this line becomes a sheet. As noted, string theories can also harbor world-volumes in the form of branes. One of the problems with a theory of such world-volumes is that internal degrees of freedom are also troublesome, and can once more give rise to consistency problems. Thus, these will be neglected here. String theory, on the other hand, seems to be singled out to be a theory with just enough smearing to avoid the problems of quantum field theory and at the same time having enough internal rigidity as to avoid new problems. Though without branes it does not appear to be able to describe nature.

One feature of string theory is that there is usually no consistent solution in four space-time dimensions, but typically more dimensions are required. How many more is actually a dynamical property of the theory: It is necessary to solve it to give an answer. In perturbation theory, it appears that ten dimensions are required, but beyond perturbation theory indications have been found that rather elven dimensions are necessary. Anyway, the number is usually larger than four. Thus, some of the dimensions have to be hidden, which can be performed by compactification, as with the setup for large extra dimensions. Indeed, as has been emphasized, large extra dimensions are rather often interpreted as a low-energy effective theory of string theory.

Since the space-time geometry of string theory is dynamic, as in case of quantum gravity, the compactification is a dynamical process. It turns out that already classically there are a huge number of (quasi-)stable solutions having a decent compactification of the surplus dimensions, but all of them harbor a different low-energy physics, i. e., a different standard model. To have the string theory choose the right vacuum, thus yielding the observed standard model, turns out to be complicated, though quantum effects actually improve the situation. Nonetheless, this problem remains a persistent challenge for string theories.

In the following the number of dimensions will be D, and the Minkowski metric will take the form

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & & 0 & \\ & 1 & & 0 \\ 0 & & \ddots & \\ & 0 & & 1 \end{pmatrix}.$$

It is actually a good question why the signature of the Minkowski metric should be like this, and also string theory so far failed to provide a convincing answer. But before turning to string theory, it makes sense to set some of the stage first with the example of a point particle.

9.5.3.1 Point particle

To become confident with the concepts take a classical particle moving along a world line in D dimensions. Classically, a trajectory is described by the $D-1$ spatial coordinates $x_i(t)$ as a function of time $t = x_0$. More useful in the context of string theory is a redundant description in terms of D functions $X_\mu(\tau)$ of a variable τ, which strictly monotonously increases along the world-line. A natural candidate for this variable is the eigentime, which thus parametrizes the world line of the particle.

The simplest Poincare-invariant action describing a free particle of mass m in terms of the eigentime is then given by

$$S_{pp} = -m \int d\tau \sqrt{-\partial_\tau X^\mu \partial^\tau X_\mu}. \tag{9.34}$$

This yields that the minimum action is obtained for the minimum (geodetic) length of the world line. Variation along the world line

$$\delta \dot{X}_\mu \equiv \delta \partial_\tau X^\mu = \partial_\tau \delta X^\mu$$

yields the equation of motion as

$$\begin{aligned}
\delta S_{pp} &= -m \int d\tau \left(\sqrt{-\dot{X}_\mu \dot{X}^\mu} - \sqrt{-\left(\dot{X}^\mu + \delta \dot{X}^\mu\right)\left(\dot{X}_\mu + \delta \dot{X}_\mu\right)} \right) \\
&= -m \int d\tau \left(\sqrt{-\dot{X}_\mu \dot{X}^\mu} - \sqrt{-\left(\dot{X}_\mu \dot{X}^\mu + 2\dot{X}^\mu \delta \dot{X}_\mu\right)} \right) \\
&= -m \int d\tau \left(\sqrt{-\dot{X}_\mu \dot{X}^\mu} - \sqrt{-\dot{X}_\mu \dot{X}^\mu \left(1 + 2\frac{\dot{X}^\mu \delta \dot{X}_\mu}{\dot{X}_\mu \dot{X}^\mu}\right)} \right) \\
&\overset{\text{Taylor}}{=} m \int d\tau \frac{\dot{X}^\mu \delta \dot{X}_\mu}{\sqrt{-\dot{X}^\mu \dot{X}_\mu}}
\end{aligned}$$

where in the last line use has been made of the infinitesimal smallness of $\delta \dot{X}_\mu$ and the square root has been Taylor-expanded.

Defining now the D-dimensional normalized speed as

$$u^\mu = \frac{\dot{X}^\mu}{\sqrt{-\dot{X}_\mu \dot{X}^\mu}} \tag{9.35}$$

yields the equation of motion after imposing the vanishing of the action under the variation and a partial integration as

$$m\dot{u}^\mu = 0 \tag{9.36}$$

This is nothing else then the equation of motion for a free relativistic particle, which of course reduces to the one of Newton in the limit of small speeds. This also justifies the interpretation of m as the rest mass of the particle.

With τ the eigentime the action is indeed Poincare-invariant. This can be seen as follows. A Poincare transformation is given by

$$X'^\mu = \Lambda^\mu_\nu X^\nu + a^\mu.$$

Inserting this expression for the argument of the square root yields

$$\partial_\tau \left(\Lambda^\mu_{\ \nu} X^\nu + a^\mu\right) \partial_\tau \left(\Lambda_\mu^{\ \omega} X_\omega + a^\mu\right) = \left(\Lambda^\mu_{\ \nu} \Lambda_\mu^{\ \omega}\right) \partial_\tau X^\nu \partial_\tau X_\omega.$$

Since the expression in parenthesis is just δ_ν^ω because of the (pseudo-)orthogonality of Lorentz transformations, this makes the expression invariant. Since the eigentime is invariant by definition, this shows the invariance of the total action.

Additionally, it is also reparametrization invariant, i. e., it is possible to transform the eigentime to a different variable without changing the contents of the theory, as it ought to be: Physics should be independent of the coordinate systems imposed by the observer. This is what ultimately leads to the diffeomorphism (diff) invariance of general relativity.

To show this invariance also for the action (9.34) take an arbitrary (but invertible) reparametrization $\tau' = f(\tau)$. This implies

$$\begin{aligned} \dot{\tau}' &= \frac{d\tau'}{d\tau} \\ d\tau &= \frac{d\tau'}{\dot{\tau}'}, \end{aligned}$$

yielding the transformation property of the integral measure. For the functions follows then

$$\dot{X}^{\mu'}(\tau') = \dot{X}^\mu(\tau)\frac{d\tau}{d\tau'} = \dot{X}^\mu \frac{1}{\dot{\tau}'}$$

Hence the scalar product changes as

$$\dot{X}^{\mu\prime}\dot{X}'_\mu = \frac{1}{\dot{\tau}'^2}\dot{X}^\mu \dot{X}_\mu.$$

One power of $\dot{\tau}'$ is removed by the square root, and the remaining one is then compensated by the integral measure.

Showing this explicitly for the action (9.34) was rather tedious, and it is useful to rewrite the action. For this purpose it is useful to introduce a metric along the world line. Since the world line is one-dimensional, this metric is only a single function $\gamma_{\tau\tau}(\tau)$ of the eigentime. This yields a trivial example of a tetrad η, which is defined as

$$\eta(\tau) = \left(-\gamma_{\tau\tau}(\tau)\right)^{\frac{1}{2}}.$$

It is in general a set of N (by definition positive) orthogonal vectors on a manifold. However, the manifold is just one-dimensional for a world line, and thus the tetrad is again only a scalar. In analogy to the string case later, $\gamma_{\tau\tau}$ can also be denoted as the world-line metric.

Taking the tetrad as an independent function a new action is defined as

$$S'_{pp} = \frac{1}{2}\int d\tau \left(\frac{\dot{X}^\mu \dot{X}_\mu}{\eta} - \eta m^2\right).$$

Under a reparametrization $\tau \to \tau'(\tau)$ it is defined that the functions X and η transform as

$$\begin{aligned} X(\tau) &= X(\tau'(\tau)) \\ \eta'(\tau') &= \eta(\tau)\frac{d\tau}{d\tau'} = \frac{1}{\dot{\tau}'}\eta(\tau) \end{aligned} \tag{9.37}$$

This makes the expression invariant under diffeomorphisms: The transformation of η (9.37) takes care of the extra factor of $\dot{\tau}'$, and also makes the second expression invariant.

To show that the new action is indeed equivalent to the old, and that η is thus just an auxiliary function, can be shown by using the equation of motion for η. Using the Euler-Lagrange equation this time yields

$$\begin{aligned} 0 = \frac{d}{d\tau}\frac{\partial L}{\partial \dot{\eta}} - \frac{\partial L}{\partial \eta} &= \frac{\dot{X}^\mu \dot{X}_\mu}{\eta^2} + m^2 \\ &\implies \eta^2 = -\frac{\dot{X}^\mu \dot{X}_\mu}{m^2}. \end{aligned}$$

Thus knowledge of X determines η completely, since no derivatives of η appear. The tetrad η is a so-called auxiliary field. Inserting this expression into (9.37) leads to

$$\begin{aligned} S'_{pp} &= \frac{1}{2}\int d\tau \left(\frac{\dot{X}^\mu \dot{X}_\mu}{\sqrt{-\frac{\dot{X}^\mu \dot{X}_\mu}{m^2}}} - \sqrt{-\frac{\dot{X}^\mu \dot{X}_\mu}{m^2}}m^2\right) \\ &= \frac{m}{2}\int d\tau \left(-\frac{-\dot{X}^\mu \dot{X}_\mu}{\sqrt{-\dot{X}^\mu \dot{X}_\mu}} - \sqrt{-\dot{X}^\mu \dot{X}_\mu}\right) \\ &= -m\int d\tau \sqrt{-\dot{X}^\mu \dot{X}_\mu} = S_{pp}. \end{aligned}$$

Thus S'_{pp} is indeed equivalent to S_{pp}. But, there is an additional advantage. By separation of the mass S'_{pp} can also be applied to the case of $m = 0$ directly, which is only possible in a limiting process for the original action S_{pp}.

9.5.3.2 Classical strings

For strings, the world line becomes a world sheet. As a consequence, at any fixed eigentime τ the string has an extension. This extension can be infinite or finite. In the latter case, the string can be closed, i. e., its ends are connected, or open. In string theories usually only finite strings appear, with lengths L of size the Planck length. Furthermore, open strings have usually to have their ends located on branes. This is not necessary for the simple case here, which will be investigated both for open and closed strings.

Analogous to the eigentime then an eigenlength σ can be introduced. Both parameters together describe any point on the world-sheet. The functions X_μ describing the position of the points of the world-sheet in space-time are therefore functions of both parameters, $X_\mu = X_\mu(\sigma,\tau)$. Furthermore, as for the point particle, these functions should be reparametrization invariant

$$X^\mu(\sigma,\tau) = X^\mu(\sigma'(\sigma,\tau),\tau'(\sigma,\tau)) \tag{9.38}$$

such that the position of the world sheet is not depending on the parametrization.

Derivatives with respect to the two parameters will be counted by Latin indices $a, ...,$

$$\begin{aligned}\partial_{a,b,...} &= \partial_\tau,\partial_\sigma\\ \partial_0 &= \partial_\tau\\ \partial_1 &= \partial_\sigma.\end{aligned}$$

It is then possible to define the induced metric on the world sheet as

$$h_{ab} = \partial_a X^\mu \partial_b X_\mu,$$

as a generalization of $h_{\tau\tau} = \dot{X}^\mu \dot{X}_\mu$, which as a metric has already been used to define the action (9.34) in analogy to the Einstein-Hilbert action. $\sqrt{-\det h_{ab}}d\tau d\sigma$ is then an infinitesimal element of the world sheet area.

The simplest possible Poincare-invariant action which can be written down for this system is the Nambu-Goto action

$$S_{NG} = \int\limits_M d\tau d\sigma \mathcal{L}_{\mathrm{NG}}$$

in which M is the world-sheet of the string and $\mathcal{L}_{\mathrm{NG}}$ is the Nambu-Goto Lagrangian

$$\mathcal{L}_{\mathrm{NG}} = -\frac{1}{2\pi\alpha'}\sqrt{-\det h_{ab}} = -\frac{1}{2\pi\alpha'}\sqrt{\partial_\tau X_\mu \partial_\sigma X^\mu \partial_\sigma X_\rho \partial_\tau X^\rho - \partial_\tau X_\mu \partial_\tau X^\mu \partial_\sigma X^\rho \partial_\sigma X_\rho},$$

again the direct generalization of the point-particle action. In particular, the minimum area of the world sheet minimizes the action.

The constant α' is the so-called Regge slope, having dimension mass squared. In principle, it could be set to one in the following for the non-interacting string, but due to its importance in the general case, it will be left explicit. The Regge slope can be associated with the string tension T as $T = 1/(2\pi\alpha')$.

The name Regge slope relates to the Regge trajectory of QCD in section 5.9, as a classically rotating string with string tension α' generates after quantization the same mass-spin-relation as is observed for Regge trajectories. Incidentally, this relation was also historically the inception of string theory, which was first invented to describe hadron physics ere QCD was discovered.

The Nambu-Goto action has two symmetries. One is diffeomorphism invariance. This can be seen directly, as in the case of the point particle, except that now the Jacobian appears. The second invariance is Poincare invariance, which leaves the world-sheet parameters τ and σ invariant. However, the functions X_μ transform as

$$\begin{aligned} X'^\mu &= \Lambda^\mu_{\ \nu} X^\nu + a^\mu \\ \partial_a \Lambda^\mu_{\ \nu} X^\nu \partial_b \Lambda_\mu^{\ \gamma} X_\gamma &= \overbrace{\Lambda^\mu_{\ \nu} \Lambda_\mu^{\ \gamma}}^{\delta_\nu^\gamma} \partial_a X^\nu \partial_b X_\gamma = \partial_a X^\mu \partial_b X_\mu . \end{aligned}$$

Thus, the induced metric is Poincare invariant, and hence also the action as well as the Lagrangian and any other quantity constructed from it is.

It is once more rather cumbersome to use an action involving a square root. To construct a simpler action, it is useful to introduce a world-sheet metric $\gamma_{ab}(\tau, \sigma)$. This metric is taken to have a Lorentz signature for some chosen coordinate system

$$\gamma_{ab} = \begin{pmatrix} + & 0 \\ 0 & - \end{pmatrix}.$$

Thus, this metric is traceless, and has a determinant smaller zero. With it the new action, the Brink-Di Vecchia-Howe-Deser-Zumino or Polyakov action,

$$S_P = -\frac{1}{4\pi\alpha'} \int\limits_M d\tau d\sigma \, (-\gamma)^{\frac{1}{2}} \, \gamma^{ab} h_{ab} \tag{9.39}$$

is constructed, where γ denotes $\det \gamma_{ab}$.

As in case of the point particle, the world-sheet metric γ_{ab} has to have a non-trivial transformation property under diffeomorphisms,

$$\frac{\partial \omega'^c}{\partial \omega^a} \frac{\partial \omega'^d}{\partial \omega^b} \gamma'_{cd}(\tau', \sigma') = \gamma_{ab}(\tau, \sigma),$$

where the variables ω denote either σ and τ, depending on the index. This guarantees that for all invertible reparameterizations, which are continuous deformations of the identity transformation, the metric is still traceless and has negative determinant.

To obtain the relation of the Polyakov action to the Nambu-Goto action it is again necessary to obtain its equation of motion. This is most conveniently obtained using the variational principle. For this, the general relation for determinants of metrics

$$\delta\gamma = \gamma\gamma^{ab}\delta\gamma_{ab} = -\gamma\gamma_{ab}\delta\gamma^{ab}$$

is quite useful.

Abbreviating the Polyakov Lagrangian by L_P and performing a variation with respect to γ yields

$$\begin{aligned}
\delta S_P &= -\frac{1}{4\pi\alpha'}\int d\tau d\sigma \left(L_P - (-\gamma-\delta\gamma)^{\frac{1}{2}}\left(\gamma^{ab}+\delta\gamma^{ab}\right)h_{ab}\right)\\
&= -\frac{1}{4\pi\alpha'}\int d\tau d\sigma \left(L_P - \left(-\gamma+\gamma\gamma^{cd}\delta\gamma_{cd}\right)^{\frac{1}{2}}\left(\gamma^{ab}+\delta\gamma^{ab}\right)h_{ab}\right)\\
&= -\frac{1}{4\pi\alpha'}\int d\tau d\sigma \left(L_P - (-\gamma)^{\frac{1}{2}}\left(1-\gamma^{cd}\delta\gamma_{cd}\right)^{\frac{1}{2}}\left(\gamma^{ab}+\delta\gamma^{ab}\right)h_{ab}\right).
\end{aligned}$$

Expanding the term with indices cd up to first order in the variation leads to

$$\begin{aligned}
\delta S_P &= -\frac{1}{4\pi\alpha'}\int d\tau d\sigma \left(L_P - (-\gamma)^{\frac{1}{2}}\left(1-\frac{1}{2}\gamma^{cd}\delta\gamma_{cd}\right)\left(\gamma^{ab}+\delta\gamma^{ab}\right)h_{ab}\right)\\
&= -\frac{1}{4\pi\alpha'}\int d\tau d\sigma \left(L_P - (-\gamma)^{\frac{1}{2}}\left(\gamma^{ab}+\delta\gamma^{ab}-\frac{1}{2}\gamma^{cd}\gamma^{ab}\delta\gamma_{cd}\right)h_{ab}\right).
\end{aligned}$$

The second term is again the Polyakov Lagrangian, canceling the zero-order term. Then only

$$\delta S_P = -\frac{1}{4\pi\alpha'}\int d\tau d\sigma \,(-\gamma)^{\frac{1}{2}}\left(h_{ab}-\frac{1}{2}\gamma_{ab}\gamma_{cd}h^{cd}\right)\delta\gamma^{ab}$$

is left.

The condition that this expression should vanish yields the equation of motion for the world-sheet metric as

$$h_{ab} = \frac{1}{2}\gamma_{ab}\gamma_{cd}h^{cd} \tag{9.40}$$

Division of each side by its determinant finally yields

$$\begin{aligned}
\frac{h_{ab}}{(-h)^{\frac{1}{2}}} &= \frac{1}{2}\frac{\gamma_{ab}\left(\gamma_{cd}h^{cd}\right)}{\left(\det -\frac{1}{2}\gamma_{ab}\gamma_{cd}h^{cd}\right)^{\frac{1}{2}}}\\
&= \frac{1}{2}\frac{\gamma_{ab}\left(\gamma_{cd}h^{cd}\right)}{\left(\left(\frac{1}{2}\gamma_{cd}h^{cd}\right)^2\det -\gamma_{ab}\right)^{\frac{1}{2}}}\\
&= \frac{\gamma_{ab}}{(-\gamma)^{\frac{1}{2}}}
\end{aligned}$$

In the second line it has been used that $\gamma_{cd}h^{cd}$ is a scalar, permitting it to pull it out of the determinant. The result implies that h and γ are essentially proportional.

Inserting this result in the Polyakov action yields

$$S_P = -\frac{1}{4\pi\alpha'}\int d\tau d\sigma \gamma^{ab}\gamma_{ab}\left(-h\right)^{\frac{1}{2}} = -\frac{1}{2\pi\alpha'}\int d\tau d\sigma \left(-h\right)^{\frac{1}{2}} = S_{NG}$$

showing that it is indeed equivalent to the Nambu-Goto action, where the fact that the diffeomorphism-invariant quantity $\gamma^{ab}\gamma_{ab}$ is two, due to the Lorentz signature of γ, has been used.

The Polyakov action thus retains the Poincare and diffeomorphism invariance of the Nambu-Goto action. The Poincare invariance follows since γ is Poincare invariant, since it is proportional to the Poincare-invariant induced metric, thus

$$\gamma^{ab\prime} = \Lambda\gamma^{ab} = \gamma^{ab}.$$

The diffeomorphism invariance follows directly from the transformation properties of the world-sheet metric, in total analogy with the point-particle case, but considerably more lengthy since track of both variables has to be kept.

The redundancy introduced with the additional degree of freedom γ grants a further symmetry. This is the so-called Weyl symmetry, given by

$$\begin{aligned} X'^{\mu}\left(\tau,\sigma\right) &= X^{\mu}\left(\tau,\sigma\right) \\ h'_{ab} &= h_{ab} \\ \gamma'_{ab} &= e^{2\omega(\tau,\sigma)}\gamma_{ab}. \end{aligned}$$

for arbitrary functions $\omega\left(\tau,\sigma\right)$. The origin of this symmetry comes from the unfixed proportionality of induced metric and the world-sheet metric. The expression of γ in terms of the induced metric h is invariant under this transformation,

$$\frac{\gamma'_{ab}}{\left(-\gamma'\right)^{\frac{1}{2}}} = \frac{\gamma_{ab}e^{2\omega}}{\left(-\gamma'\right)^{\frac{1}{2}}} = \frac{\gamma_{ab}e^{2\omega}}{\left(-\gamma e^{4\omega}\right)^{\frac{1}{2}}} = \frac{\gamma_{ab}}{\left(-\gamma\right)^{\frac{1}{2}}}.$$

Also the action is invariant. To see this note that γ_{ab} is indeed a metric. Since $\gamma_{ab}\gamma^{ab}$ has to be a constant, as noted before, this implies that

$$\gamma'^{ab} = e^{-2\omega}\gamma^{ab}.$$

As a consequence, the expression appearing in the action transforms as

$$\left(-\gamma'\right)^{\frac{1}{2}}\gamma'^{ab} = \left(-\gamma e^{4\omega}\right)^{\frac{1}{2}}\gamma^{ab}e^{-2\omega} = \left(-\gamma\right)^{\frac{1}{2}}\gamma^{ab}.$$

Thus, the Weyl invariance is indeed a symmetry.

The Polyakov action can also be viewed with a different interpretation. Promoting the world-sheet indices to space-time indices and taking the indices μ to label internal degrees of freedom, then the Polyakov action just describes D massless Klein-Gordon fields X_μ (with internal symmetry group SO($D-1$,1)) in two space-time dimensions with a non-trivial metric γ, which is dynamically coupled to the fields.

9.5.3.3 Light cone gauge

As the Poincare and Weyl symmetry introduce a gauge symmetry, it is easier[17] to perform the quantization in a fixed gauge. Particularly useful for this purpose in the present context is the light-cone gauge. Though this gauge is not keeping manifest Poincare covariance, it is very useful (similar to the case of quantizing electrodynamics in Coulomb rather than Landau gauge). Proving that the theory is still covariant after quantization is non-trivial, but possible. Hence, this will not be shown here.

To formulate light-cone gauge light-cone coordinates are useful. They are introduced by the definitions

$$\begin{aligned} x^{\pm} &= \frac{1}{\sqrt{2}}\left(x^0 \pm x^1\right) \\ x^i &= x^i,\ i=2,...,D-1, \end{aligned}$$

and thus mix the time-coordinate and one, now distinguished, spatial coordinate. Since the zero-component is the only one involving a non-positive sign in the metric this yields the following relation between covariant and contravariant light-cone coordinates

$$\begin{aligned} x_{\pm} &= \frac{1}{\sqrt{2}}\left(x_0 \mp x_1\right) \\ x_- &= -x^+ \\ x_+ &= -x^- \\ x_i &= x^i. \end{aligned}$$

This implies the metric

$$a^\mu b_\mu = a^+b_+ + a^-b_- + a^ib_i = -a^+b^- - a^-b^+ + a^ib^i,$$

which is equivalent to the conventional one

$$\begin{aligned} -a^+b^- - a^-b^+ + a^ib^i &= -\frac{1}{2}\left(a^0+a^1\right)\left(b^0-b^1\right) - \frac{1}{2}\left(a^0-a^1\right)\left(b^0+b^1\right) + a^ib^i \\ &= -a^0b^0 + a^1b^1 + a^ib^i = a^\mu b_\mu, \end{aligned}$$

[17] Actually, in a continuum theory this may even be necessary.

as expected.

Aim of the gauge fixing is to restore the original number of independent degrees of freedom. In case of the point particle this amounts to remove the eigentime τ. This is most conveniently done by the condition

$$\tau = x^+,$$

thus being the light-cone gauge condition for the point particle. This is more convenient than the more conventional choice $\tau = x^0$. With this x^+ corresponds to the time and p^- to the energy . Correspondingly, x^- and p^+ are now longitudinal degrees of freedom while x^i and p^i are transverse ones. This immediately follows from the scalar product

$$\frac{\partial}{\partial a^+}\left(-a^+ b^- + ...\right) = -b^-,$$

and correspondingly for the derivative with respect to x^+ which produces p^-.

9.5.3.4 Point particle: Quantization

To demonstrate the principles, it is once more convenient to first investigate the point particle. However, one should be warned that the resulting theory is actually flawed due to the appearance of unphysical (non-normalizable) states. It should therefore be taken rather as a mathematical than a physical discussion.

Returning to the parametrization of the point particle of section 9.5.3.1, the gauge condition to fix the diffeomorphism invariance becomes

$$X^+(\tau) = \tau.$$

The action is then already given by equation (9.37), thus

$$\begin{aligned} S'_{pp} &= \frac{1}{2}\int d\tau \left(\frac{\dot{X}^\mu \dot{X}_\mu}{\eta} - \eta m^2\right) \\ &= \frac{1}{2}\int d\tau \left(\frac{1}{\eta}\left(-\dot{X}^+\dot{X}^- - \dot{X}^-\dot{X}^+ + \dot{X}^i\dot{X}^i\right) - \eta^2 m\right) \\ &= \frac{1}{2}\int d\tau \left(\frac{1}{\eta}\left(-2\dot{X}^-\dot{\tau} + \dot{X}^i\dot{X}^i\right) - \eta^2 m\right) \\ &= \frac{1}{2}\int d\tau \left(\frac{1}{\eta}\left(-2\dot{X}^- + \dot{X}^i\dot{X}^i\right) - \eta m^2\right). \end{aligned}$$

As usual, the Lagrangian yields the canonical conjugated momenta by the expression

$$P_\mu = \frac{\partial L}{\partial \dot{X}^\mu}$$

yielding

$$\begin{aligned} P_- &= -\frac{1}{\eta} \\ P_i &= \frac{\dot{X}^i}{\eta} \end{aligned}$$

With this the Hamiltonian can be readily constructed as

$$\begin{aligned} H &= \sum P\dot{Q} - L \\ &= P_-\dot{X}^- + P_i\dot{X}^i - L \\ &= -\frac{\dot{X}^-}{\eta} + \eta P_i P_i + \frac{\dot{X}^-}{\eta} - \frac{1}{2}\dot{X}^i\dot{X}^i + \frac{1}{2}\eta m^2 \\ &= \eta P_i P_i - \frac{1}{2}\eta P_i P_i + \frac{1}{2}\eta m^2 = \frac{P^i P^i + m^2}{2P^+}. \end{aligned}$$

Where it has been used that (9.41) implies

$$P^+ = -P_- = \frac{1}{\eta},$$

and it is thus possible to remove η and P_- from the expression.

In this result the variable X^+ is no longer a dynamical variable, and thus the gauge is fixed. Furthermore it follows that $P_\eta = 0$, since the Lagrangian does not depend on $\dot{\eta}$. Hence η is not a dynamical variable. This was expected, since it was already in the classical case only used to make the Lagrangian more easily tractable

For the quantization then the usual canonical commutation relations are imposed as

$$\begin{aligned} \left[P_i, X^j\right] &= -i\delta_i^j \\ \left[P_-, X^-\right] &= -i \end{aligned}$$

The relations for P_+ is provided by the other relations, since P^- is the energy and thus

$$H = P^- = -P_+. \tag{9.41}$$

That is essentially the relativistic mass-shell equation, implying once more that P^+ is not an independent degree of freedom. The resulting Hamilton operator is the one of a $D-2$-dimensional harmonic oscillator, but supplemented with the additional unconstrained degree of freedom P_-. The spectrum of this is known, being a relativistic scalar (with all its sicknesses) and states $|k_-, k_i\rangle$.

9.5.3.5 Open string

Again, the first step is to fix the gauge. For that purpose first the permitted range for the world-sheet parameters have to be chosen, which will be

$$\begin{aligned} -\infty &\leq \tau \leq +\infty \\ 0 &\leq \sigma \leq L. \end{aligned} \tag{9.42}$$

Thus, L is the length of the string. Again, it is chosen that

$$\tau = X^+. \tag{9.43}$$

This deals again with the diffeomorphism degree of freedom. To also take care of the Weyl freedom a second condition is necessary, which will be chosen to be

$$\partial_\sigma \gamma_{\sigma\sigma} = 0 \tag{9.44}$$

$$\det \gamma_{ab} = -1 \tag{9.45}$$

The conditions (9.43) and (9.44-9.45) fixes these degrees of freedom completely, provided that the world-sheet is parametrized by the eigenvariables in such a way that one and only one set of eigentime and eigenlength correspond to a given point on the world sheet. In the case of the point particle, it can be shown that this condition is actually superfluous, since even in case of a doublebacking world line this would not contribute to a path integral. For string theory, this is something not yet really simply understood.

A way to get an intuition for the significance of these gauge condition is by the use of the invariant length. The choice of $\tau = X^+$ is of course always possible. Then start by the definition

$$f = \gamma_{\sigma\sigma}\left(\frac{1}{-\det \gamma_{ab}}\right)^{\frac{1}{2}}.$$

Now perform a reparametrization which leaves τ invariant. This implies

$$f' = f\frac{d\sigma}{d\sigma'}.$$

because of the transformation properties of the γ_{ab}. Hence, the length element $dl = f d\sigma$ is invariant under this reparametrization. Therefore, it can be considered as an invariant length-element, since it is not changing under a change of the eigenlength of the string. In fact, this can be used to define the σ coordinate, by setting it equal to $\int dl$ along the world sheet,

$$\sigma = \int_0^\sigma dl.$$

As a consequence, f can no longer depend on σ, since dl is σ-independent. Secondly, it is then possible to make a Weyl-transformation to rescale $\det\gamma$ such that it becomes -1, yielding (9.45), and fixing the Weyl invariance. Since f is Weyl-invariant by construction, this implies that $\partial_\sigma\gamma_{\sigma\sigma}$ trivially vanishes, yielding (9.44). Thus, in this coordinate system the gauge condition are fulfilled, and therefore are a permitted choice.

Since γ is by construction symmetric, these gauge conditions permit to rewrite it in a simpler way. It then takes the form

$$
\begin{aligned}
\gamma &= \begin{pmatrix} \gamma^{\tau\tau} & \gamma^{\tau\sigma} \\ \gamma^{\sigma\tau} & \gamma^{\sigma\sigma} \end{pmatrix} \\
&= \begin{pmatrix} -\gamma_{\sigma\sigma}(\tau) & \gamma_{\tau\sigma}(\tau,\sigma) \\ \gamma_{\tau\sigma}(\tau,\sigma) & \gamma_{\sigma\sigma}^{-1}(\tau)\left(1-\gamma_{\tau\sigma}^2(\tau,\sigma)\right) \end{pmatrix},
\end{aligned}
$$

thereby eliminating two of the four variables in γ_{ab}, and also reducing their dependence on the world sheet parameters.

It is furthermore useful to define the average and variation of the X^- coordinate for the following as

$$
\begin{aligned}
Z^-(\tau) &= \frac{1}{L}\int\limits_0^L d\sigma X^-(\tau,\sigma) \\
Y^-(\tau,\sigma) &= X^-(\tau,\sigma) - Z^-(\tau).
\end{aligned}
$$

This is the starting point to rewrite the action in a more useful form.

Start by rewriting the Lagrangian as

$$
\begin{aligned}
L_P &= -\frac{1}{4\pi\alpha'}\int\limits_0^L d\sigma \overbrace{(-\gamma)^{\frac{1}{2}}}^{=1} \gamma^{ab}\partial_a X^\mu \partial_b X_\mu \\
&= -\frac{1}{4\pi\alpha'}\int\limits_0^L \left(\frac{1-\gamma_{\tau\sigma}^2}{\gamma_{\sigma\sigma}}\partial_\sigma X^\mu\partial_\sigma X_\mu + \gamma_{\tau\sigma}\partial_\tau X^\mu \partial_\sigma X_\mu + \gamma_{\tau\sigma}\partial_\sigma X^\mu\partial_\tau X_\mu - \gamma_{\sigma\sigma}\partial_\tau X^\mu \partial_\tau X_\mu\right).
\end{aligned}
$$

Now, it is useful to investigate the expressions piece-by-piece. Start with

$$
\frac{1-\gamma_{\tau\sigma}^2}{\gamma_{\sigma\sigma}}\partial_\sigma X^\mu\partial_\sigma X_\mu = \frac{1-\gamma_{\tau\sigma}^2}{\gamma_{\sigma\sigma}}\partial_\sigma X^i\partial_\sigma X^i
$$

where it has been used that

$$
\partial_\sigma X^+ = \partial_\sigma\tau = 0
$$

follows trivially from the gauge conditions. Next, use furthermore that

$$\gamma_{\tau\sigma}(\overbrace{\partial_\tau X^+\partial_\sigma X_+}^{=-\partial_\tau\tau\partial_\sigma X^-}+\partial_\tau X^-\overbrace{\partial_\sigma X_-}^{=-\partial_\sigma X^+}+\partial_\tau X^i\partial_\sigma X^i)=\gamma_{\tau\sigma}\left(-\partial_\sigma X^-+\partial_\sigma X^i\partial_\tau X^i\right)$$

and

$$-\gamma_{\sigma\sigma}\left(\partial_\tau X^+\partial_\tau X_++\partial_\tau X^-\partial_\tau X_-+\partial_\tau X^i\partial_\tau X^i\right)=-\gamma_{\sigma\sigma}\left(-2\partial_\tau X^-+\partial_\tau X^i\partial_\tau X^i\right).$$

Reinserting everything into the Lagrangian yields

$$\begin{aligned} L_P &= -\frac{1}{4\pi\alpha'}\int\limits_0^L d\sigma\Big(\frac{1-\gamma_{\tau\sigma}^2}{\gamma_{\sigma\sigma}}\partial_\sigma X^i\partial_\sigma X^i-2\gamma_{\tau\sigma}\left(-\partial_\sigma X^-+\partial_\tau X^i\partial_\sigma X^i\right)\\ &\quad -\gamma_{\sigma\sigma}\left(-2\partial_\tau X^-+\partial_\tau X^i\partial_\tau X^i\right)\Big).\end{aligned}$$

Employing now the relations for the average and variation this yields

$$\begin{aligned} L_P &= -\frac{1}{4\pi\alpha'}\Big(\gamma_{\sigma\sigma}2L\partial_\tau Z^-+\int\limits_0^L d\sigma\Big(\gamma_{\sigma\sigma}\left(-\partial_\tau X^i\partial_\tau X^i\right)\\ &\quad +2\gamma_{\tau\sigma}\left(\partial_\sigma Y^--\partial_\tau X^i\partial_\sigma X^i\right)+\frac{1-\gamma_{\tau\sigma}^2}{\gamma_{\sigma\sigma}}\partial_\sigma X^i\partial_\sigma X^i\Big)\Big)\end{aligned}$$

In the resulting expression there is no τ-derivative of Y^- appearing, which is thus a non-dynamical field, behaving like a Lagrange factor for $\gamma_{\tau\sigma}$, which therefore is fixed to

$$\partial_\sigma\gamma_{\tau\sigma}=0, \tag{9.46}$$

and thus does not depend on σ.

Returning to the boundary condition of this open[18] string yields

$$\left(\partial_\sigma X^\mu\right)(\tau,0)=\left(\partial_\sigma X^\mu\right)(\tau,L)=0, \tag{9.47}$$

because otherwise the fields would not be continuously differentiable at the boundaries, which is imposed like for wave-functions. These are Neumann conditions in the terms of the large extra dimensions. This is also obtained by varying the Polyakov action. First, vary with respect to the fields to obtain

$$-\frac{1}{2\pi\alpha'}\int\limits_{-\infty}^{\infty}d\tau\,(-\gamma)^{\frac{1}{2}}\,\partial_\sigma X^\mu\delta X^\mu|_{\sigma=0}^{\sigma=L}.$$

[18]No cyclicity of any function on σ has been imposed, which would be one possibility to implement a closed string.

Since this has to vanish for arbitrary variations of the fields, this implies the boundary condition (9.47).

On the other hand, when varying the original action with respect to the fields, this yields

$$\begin{aligned}\delta S_P &= S_P + \frac{1}{4\pi\alpha'}\int d\tau d\sigma \gamma_{ab}\partial_a\left(X^\mu+\delta X^\mu\right)\partial_b\left(X_\mu+\delta X_\mu\right)\\ &= \frac{1}{4\pi\alpha'}\int d\tau d\sigma\gamma_{ab}\left(\partial_a X^\mu\partial_b\delta X_\mu+\partial_\alpha\delta X^\mu\partial_b X_\mu\right).\end{aligned}$$

Since variation and differentiation are independent, they can be exchanged,

$$\partial_a\delta X^\mu = \delta\partial_a X^\mu.$$

Doing a partial integration, keeping an appearing boundary term yields

$$\begin{aligned}\delta S_P &= \frac{1}{4\pi\alpha'}\int d\tau d\sigma\gamma_{ab}\left(\partial_a\partial_b X^\mu\delta X_\mu+\partial_b\partial_a X_\mu\delta X^\mu\right)\\ &\quad -\frac{1}{4\pi\alpha'}\int d\tau\gamma_{ab}\left(\partial_a X^\mu\delta X_\mu+\partial_b X_\mu\delta X^\mu\right)\Big|_0^L\end{aligned}\tag{9.48}$$

Note that in the boundary term as a shorthand notation one of the indices is uncontracted. This is of course always the σ-index for which the total integration has been performed. However, the last expression must vanish under variation, implying once more the Neumann condition (9.47)

$$\gamma_{ab}\left(\partial_a X^\mu+\partial_b X^\mu\right)=0.$$

Incidentally, this also implies for $\mu = +$ and $a=\tau$ and $b=\sigma$ that $\gamma_{\tau\sigma}$ vanishes on the boundary.

Since for $\mu = -$ the fields are non-dynamical, this implies that $\partial_\sigma X^- = 0$ and that therefore X^- only depends on τ.

To obtain some further useful results, the variation can be repeated after the gauge has been fixed. This yields

$$\begin{aligned}\delta S_P = & S_P + \frac{1}{4\pi\alpha'}\int d\tau d\sigma\Big(\gamma_{\sigma\sigma}\left(2\partial_\tau\left(X^-+\delta X^-\right)-\partial_\tau\left(X^i+\delta X^i\right)\partial_\tau\left(X^i+\delta X^i\right)\right)\\ & +2\gamma_{\tau\sigma}\left(\partial_\sigma\left(X^-+\delta X^-\right)-\partial_\tau\left(X^i+\delta X^i\right)\partial_\sigma\left(X^i+\delta X^i\right)\right)\\ & +\frac{1-\gamma_{\tau\sigma}^2}{\gamma_{\sigma\sigma}}\partial_\sigma\left(X^i+\delta X^i\right)\partial_\sigma\left(X^i+\delta X^i\right)\Big).\end{aligned}$$

Expanding the result and dropping $\mathcal{O}\left(\delta^2\right)$ terms and annihilating a term of type S_P just leaves

$$\begin{aligned}\delta S_P = & \frac{1}{4\pi\alpha'}\int d\tau d\sigma\Big(\gamma_{\sigma\sigma}\left(2\partial_\tau\delta X^- - 2\partial_\tau X^i\partial_\tau\delta X^i\right)\\ & +2\gamma_{\tau\sigma}\left(\partial_\sigma\delta X^- - \partial_\tau X^i\partial_\sigma\delta X^i-\partial_\tau\delta X^i\partial_\sigma X^i\right)+2\frac{1-\gamma_{\tau\sigma}^2}{\gamma_{\sigma\sigma}}\partial_\sigma X^i\partial_\sigma\delta X^i\Big)\end{aligned}$$

which can be rewritten as

$$\begin{aligned} S_P &= \frac{1}{4\pi\alpha'}\int d\tau d\sigma\Big(\left(2\gamma_{\tau\sigma}\partial_\sigma\delta X^-\right)+\left(-2\gamma_{\tau\sigma}\partial_\tau X^i\partial_\sigma\delta X^i+2\frac{1-\gamma_{\tau\sigma}^2}{\gamma_{\sigma\sigma}}\partial_\sigma\delta X^i\partial_\sigma X\right) \\ &+\left(\gamma_{\sigma\sigma}2\partial_\tau\delta X^- -2\gamma_{\sigma\sigma}\partial_\tau X^i\partial_\tau\delta X^i\right)-\left(\gamma_{\tau\sigma}\partial_\tau\delta X^i\partial_\sigma X^i\right)\Big). \end{aligned}$$

After partial integration of the first term this yields once more that $\partial_\sigma\gamma_{\tau\sigma}$ still vanishes at the end of the string.

The second term in parentheses yields after partial integration

$$-\partial_\sigma\left(2\gamma_{\tau\sigma}\partial_\tau X^i\right)+\partial_\sigma\left(\frac{1-\gamma_{\tau\sigma}^2}{\gamma_{\sigma\sigma}}\partial_\sigma X^i\right)=-2\gamma_{\tau\sigma}\partial_\sigma\partial_\tau X^i+\frac{1-\gamma_{\tau\sigma}^2}{\gamma_{\sigma\sigma}}\partial_\sigma^2 X^i.$$

Again, this boundary term has to vanish. The second does this, if the derivative of the X^i with respect to σ does so at the boundary, again yielding (9.47). Since this is not the case for the τ-derivative, this again requires $\gamma_{\tau\sigma}=0$ at the boundary of the string. Hence, this implies that both the function and its first derivative vanishes on the boundary. Because of the equation of motion for $\gamma_{\tau\sigma}$ (9.46), this implies

$$\gamma_{\tau\sigma}\equiv 0,$$

and it can be dropped everywhere.

This eliminates one degree of freedom, leaving only

$$Z^-(\tau),\ \gamma_{\sigma\sigma}(\tau),\ X^i(\tau,\sigma),$$

which is a rather short list. Furthermore, this simplifies the Polyakov Lagrangian to

$$L_P=-\frac{L}{2\pi\alpha'}\gamma_{\sigma\sigma}\partial_\tau Z^-+\frac{1}{4\pi\alpha'}\int_0^L d\sigma\left(\gamma_{\sigma\sigma}\partial_\tau X^i\partial_\tau X^i-\frac{1}{\gamma_{\sigma\sigma}}\partial_\sigma X^i\partial_\sigma X^i\right),$$

which will now serve as the starting point for quantization. It should be noted that the gauge-fixing was the reason for eliminating the degrees of freedom, reducing the set to a one more manageable for the following.

The first step for quantization is then the calculation of the canonical momenta

$$\begin{aligned} P_- &= -P^+=\frac{\partial L_P}{\partial\left(\partial_\tau Z^-\right)}=-\frac{L}{2\pi\alpha'}\gamma_{\sigma\sigma} \\ \Pi^i &= \frac{\delta L_P}{\delta\left(\partial_\tau X^i\right)}=\frac{1}{2\pi\alpha'}\gamma_{\sigma\sigma}\partial_\tau X^i=\frac{P^+}{L}\partial_\tau X^i. \end{aligned} \tag{9.49}$$

From this the Hamiltonian is immediately constructed to be

$$\begin{aligned} H &= P_{-}\partial_\tau Z^- + \int_0^L d\sigma \Pi^i \partial_\tau X^i - L \\ &= \frac{L}{4\pi\alpha' P^+} \int_0^L d\sigma \left(2\pi\alpha' \Pi^i \Pi^i + \frac{1}{2\pi\alpha'} \partial_\sigma X^i \partial_\sigma X^i \right) \end{aligned} \tag{9.50}$$

This is the Hamiltonian for $D-2$ free fields X^i and the conserved quantity P^+, as can be seen form the equations of motion

$$\begin{aligned} \partial_\tau Z^- &= \frac{\partial H}{\partial P^-} = \frac{H}{P^+} \\ \partial_\tau P^+ &= -\frac{\partial H}{\partial Z^-} = 0 \\ \partial_\tau X^i &= \frac{\delta H}{\delta \Pi^i} = 2\pi\alpha' c \Pi^i & (9.51) \\ \partial_\tau \Pi^i &= -\frac{\delta H}{\delta X^i} = \frac{c}{2\pi\alpha'} \partial_\sigma^2 X^i, & (9.52) \end{aligned}$$

where a partial integration has been performed in (9.52) and c is defined as

$$c := \frac{L}{2\pi\alpha' P^+}.$$

Inserting (9.51) in (9.52) yields the wave equation for X^i

$$\partial_\tau^2 X^i = c^2 \partial_\sigma^2 X^i,$$

where c takes the role of the wave speed. Thus, the transverse degrees of freedom form waves along the string.

Since P^+ and L are constants of motion, so is c. Thus, given the boundary conditions for the open string, the equations of motions can be solved, yielding

$$\begin{aligned} \hat{X}^i(\tau,\sigma) &= \hat{Z}^i + \frac{\hat{P}^i}{P^+}\tau + i(2\alpha')^{\frac{1}{2}} \sum_{n=-\infty n\neq 0}^{n=\infty} \frac{\alpha_n^i}{n} e^{-\frac{\pi i n c \tau}{L}} \cos\frac{\pi n \sigma}{L} & (9.53) \\ \alpha_{-n}^i &= \alpha_n^{i+}. & (9.54) \end{aligned}$$

The relation (9.54) applies since the X^i are real functions. For the purpose at hand also the center-of-mass variables

$$\begin{aligned} \hat{Z}^i(\tau) &= \frac{1}{L} \int_0^L d\sigma \hat{X}^i(\tau,\sigma) \\ \hat{P}^i(\tau) &= \int_0^L d\sigma \Pi^i(\tau,\sigma) = \frac{P^+}{L} \int_0^L d\sigma \partial_\tau X^i(\tau,\sigma) \end{aligned}$$

have been introduced. Thus, the center-of-mass of the string follows a free, linear trajectory in space, which overlays the transverse motions of the oscillations transverse to the string. Herein $\hat{Z}^i$ and $\hat{P}^i$ in (9.53) have to be taken at $\tau = 0$, and will become Schrödinger operators in the quantization procedure to come now.

The quantization procedure is started as usually with imposing equal-time canonical commutation relations

$$\begin{aligned} \left[Z^-, P^+\right] &= -i \\ \left[X^i(\sigma), \Pi^j(\sigma')\right] &= i\delta^{ij}\delta(\sigma - \sigma') \end{aligned}$$

Performing a Fourier expansion this is equivalent to the relations

$$\begin{aligned} \left[\hat{X}^i, \hat{P}^j\right] &= i\delta^{ij} \\ \left[\alpha^i_m, \alpha^j_n\right] &= m\delta^{ij}\delta_{m,-n} \end{aligned} \tag{9.55}$$

Here, a non-standard, though useful, normalization of (9.55) has been performed.

The natural consequence is now that every transverse component behaves as a harmonic oscillator with a non-standard normalization. The corresponding creation and annihilation operators are then given for $m > 0$

$$\begin{aligned} \alpha^i_m &= \hbar\sqrt{m}a \\ \alpha^i_{-m} &= \hbar\sqrt{m}a^\dagger \\ -1 &= \left[a^+, a\right] \end{aligned} \tag{9.56}$$

where m gives the oscillator level for direction i. So far, so standard.

Defining now the momentum vector $k = (k^+, k^i)$ the state $|0, k\rangle$ of lowest excitation has the properties

$$\begin{aligned} P^+ |0, k\rangle &= k^+ |0, k\rangle \\ P^i |0, k\rangle &= k^i |0, k\rangle \\ \alpha^i_m |0, k\rangle &= 0 \text{ for } m > 0 \end{aligned} \tag{9.57}$$

Therefore k is the center-of-mass momentum. Higher excited states are then denoted by $|N, k\rangle$ and can be constructed as

$$|N, k\rangle = \left(\prod_{i=2}^{D-1}\prod_{n=1}^{\infty} \frac{\left(\alpha^i_{-n}\right)^{N_{in}}}{\sqrt{(n^{N_{in}}N_{in})!}}\right)|0, k\rangle,$$

just as ordinary oscillator states. Therefore, N_{in} are the occupation numbers for each direction and level. In particular, these can be interpreted as internal degrees of freedom, while the motion of the center-of-mass corresponds to a particle like behavior of the whole string. As will be discussed below, from this point of view every state corresponds to a certain particle with a certain spin.

The total set of states (9.53) forms the Hilbert space of a single string, H_1. In particular, $|0,0\rangle$ is not a vacuum, but merely a momentum-zero string with no internal excitations, except zero-point oscillations: A quantum-mechanical string always quivers. The vacuum is devoid of a string, its Hilbert-space H_0 is denoted by the single state $|\text{vac}\rangle$. However, none of the operators so far can mediate between H_0 and H_1, but only act inside H_1. Since there are no interactions, an N-string Hilbert space can be build just as a product space of H_1s as

$$h_n = |\text{vac}\rangle \oplus H_1 \oplus ... \oplus H_n.$$

where the states are implicitly symmetrized, since the string states are bosonic, given that their creation and annihilation operators fulfill bosonic canonical commutation relations, (9.56).

Since the states are just free states, it is straightforward to construct the number-state version of the Hamiltonian. For this purpose, it is necessary to calculate the explicit form of the canonical momentum operators Π^i first as

$$\begin{aligned} \Pi^i &= \frac{P^+}{L}\left(\partial_\tau X^i\right) \\ &= \frac{P^+}{L}\left(\frac{\hat{P}^i}{P^+} + \frac{\pi c}{L}\left(2\alpha'\right)^{\frac{1}{2}} \sum_{n=-\infty n\neq 0}^{n=+\infty} \alpha_n^i e^{-\frac{\pi i n c\tau}{L}} \cos\frac{\pi n\sigma}{L}\right). \end{aligned}$$

In addition, also $\partial_\sigma X^i$ is required, and is given by

$$\partial_\sigma X^i = -\frac{i\pi}{L}\left(2\alpha'\right)^{\frac{1}{2}} \sum_{n=-\infty n\neq 0}^{n=+\infty} \alpha_n^i e^{-\frac{\pi i n c\tau}{L}} \sin\frac{\pi n\sigma}{L}.$$

Putting everything together yields the Hamiltonian

$$\begin{aligned} &\frac{L}{4\pi\alpha' P^+}\int_0^L d\sigma\left(2\pi\alpha\Pi^i\Pi^i + \frac{1}{2\pi\alpha'}\partial_\sigma X^i\partial_\sigma X^i\right) \\ = &\frac{L}{4\pi\alpha' P^+}\Bigg(2\pi\alpha' P^i P^i + \int_0^L d\sigma \\ &\Bigg(\frac{\pi}{4\alpha' L P^+}\sum_{n=-\infty n\neq 0}^{n=+\infty} \alpha_n^i e^{-\frac{\pi i n c\tau}{L}} \cos\frac{\pi n\sigma}{L} \sum_{m=-\infty m\neq 0}^{m=+\infty} \alpha_m^i e^{-\frac{\pi i m c\tau}{L}} \cos\frac{\pi m\sigma}{L} \\ &-\frac{\pi}{4\alpha' L P^+}\sum_{n=-\infty n\neq 0}^{n=+\infty} \alpha_n^i e^{-\frac{\pi i n c\tau}{L}} \sin\frac{\pi n\sigma}{L} \sum_{m=-\infty m\neq 0}^{m=+\infty} \alpha_m^i e^{-\frac{\pi i m c\tau}{L}} \sin\frac{\pi m\sigma}{L}\Bigg)\Bigg), \end{aligned}$$

where in the integration σ was replaced by $\pi\sigma/L$. Since sine and cosine are orthogonal, the integrations can be performed explicitly. Those over cos yield $\pi\delta_{n-m}$, while those over sin yield $-\pi\delta_{n-m}$. This leads to

$$H = \frac{P^i P^i}{2P^+} + \frac{1}{2P^+\alpha'} \sum_{n=-\infty n\neq 0}^{n=+\infty} \alpha_n^i \alpha_{-n}^i,$$

and finally by rearranging to

$$H = \frac{P^i P^i}{2P^+} + \frac{1}{2P^+\alpha'} \sum_{n=1}^{\infty} \alpha_{-n}^i \alpha_n^i + A.$$

This Hamiltonian is already in so-called normal order, i. e. annihilation operators to the right of all creation operators, and A is a (divergent) constant which appears in the process of normal ordering.

The actual value of A can be determined by explicitly verifying the Lorentz covariance of the result, since the Hamiltonian is just the energy, and thus a zero-component of a four-vector. However, in light-cone gauge this is far from trivial, and this will therefore be done here only in a rather sketchy way.

First, consider the zero-point energy. Every oscillator will have a zero-point energy of $\omega/2 = 1/(2P^+\alpha')$, while the transverse momenta P^i will be 0. In total, at zero excitation, it should be expected that

$$\langle 0,0|H|0,0\rangle = A,$$

due to the normal ordering. Due to the non-standard normalization, each oscillator actually contributes $n\omega/2$ of vacuum energy to this value. These oscillations appear for $D-2$ dimensions. Rewriting A as ωA this yields[19]

$$A = \frac{D-2}{2} \sum_{n=1}^{\infty} n,$$

which is, of course, infinite. However, in contrast to normal quantum mechanics or quantum field theory, the vacuum energy is not necessarily a scalar, but may couple to gravity. It is therefore necessary to maintain Lorentz invariance when treating it, and it cannot be absorbed just in a redefinition of the zero-point energy, as in quantum mechanics.

To regularize the result Lorentz-invariantly, it is necessary to include a cut-off function

$$e^{-\frac{\varepsilon|k_\sigma|}{\sqrt{\gamma_{\sigma\sigma}}}}$$

$$k_\sigma = \frac{n\pi}{L},$$

[19]With standard normalization, n would be replaced by 1, changing nothing qualitatively.

in the sum, and taking the limit $\varepsilon \to 0$ only after summation. The factor of $\sqrt{\gamma_{\sigma\sigma}}^{-1}$ is required to maintain the effects of reparametrization invariance correctly. The reason for this is that, outside light-cone gauge, the string length is not fixed, but can be changed by a reparametrization. Therefore, k_σ, which depends on the length of the string, changes under such transformations. Including the function of $\gamma_{\sigma\sigma}$ exactly cancels this effect.

Inserting this expression into the sum permits to evaluate it exactly, yielding

$$\begin{aligned} A &= \frac{D-2}{2}\sum_{n=1}^{\infty} n e^{-\frac{\varepsilon|k_\sigma|}{\sqrt{\gamma_{\sigma\sigma}}}} \\ &= \frac{D-2}{2}\left(\frac{2LP^+\alpha'}{\varepsilon^2\pi} - \frac{1}{12} + O(\varepsilon)\right), \end{aligned}$$

where the second line of (9.49) has been used . The first term is proportional to L and can therefore be absorbed in the action by an additional term proportional to

$$-\int d^2\sigma\,(-\gamma)^{\frac{1}{2}} = -\int d\sigma L.$$

This is a constant, and therefore is not changing the action. In fact, the value of the action has to be regularized itself by a similar expression, and also regularized by $e^{-\varepsilon}$. Thus, by appropriately selecting the pre-factors, both terms cancel. Since also the last term vanishes in the limit of $\varepsilon \to 0$, the only thing remaining is

$$A = \frac{2-D}{24}, \tag{9.58}$$

which is known as the Casimir energy, and can be traced back to the fact that the string is only of finite length. Thus, the string has indeed a non-zero vacuum energy. In contrast to the first contribution, this constant, non-divergent term cannot be naturally absorbed by a counter term in the action without spoiling Lorentz invariance.

Having now obtained the Hamiltonian and the state space, it is about time to determine the properties of the physical state space. In particular, the question is whether the string excitations can be interpreted as particle states, the original motivation to study it. For that purpose the primary object is of course whether the states satisfy the energy-momentum relation of a point particle, and if yes, what are their masses.

The corresponding operator for the rest mass is just given by the mass-shell equation, where it is to be used that $P^- = H$ to yield

$$m^2 = 2P^+H - P^iP^i, \tag{9.59}$$

as a result of the light-cone equation

$$m^2 = P^+P^- + P^-P^+ - P^iP^i.$$

Inserting the result (9.58) into (9.59) for the lowest-energy state yields

$$\begin{aligned} m^2 &= 2P^+\left(\frac{P^iP^i}{2P^+}+\frac{1}{2P^+\alpha'}(N+A)\right)-P^iP^i \\ &= \frac{1}{\alpha'}\left(N+\frac{2-D}{24}\right). \end{aligned}$$

That is quite an important result, as it implies that the mass is only dependent on the state sum N defined as

$$N=\sum_{i=2}^{D-1}\sum_{n=1}^{\infty}nN_{in}$$

and the space-time dimensionality D. Thus, mass becomes an intrinsic property rather than an external parameter as in the standard model. The importance of the Regge slope is now also clear, as it links as constant of proportionality the number of a state and its rest mass.

The lowest state is of course $N=0$, hence $|0,k\rangle$, and this yields

$$m^2=\frac{2-D}{24\alpha'}$$

Since for any phenomenologically relevant string theory $D>2$ the rest mass of the lowest state is imaginary, $m^2<0$. Thus it is a tachyon. That is of course unfortunate, since interpreting this as a particle is very problematic. E. g., when constructing a theory of such non-interacting scalar tachyons this yields a potential energy proportional to $m^2\phi^2/2$. Hence, the vacuum state is unstable. Of course, this would be the lowest approximation, and it could still be that the bosonic string theory is nonetheless stable, but this is unknown so far. Fortunately, in particular in supersymmetric string theories tachyons usually do not appear, so they provide a possibility to circumvent this problem without having to deal with it explicitly.

The first non-tachyonic state is obtained for the state $\alpha_{-1}^i|0,k\rangle$ with $N=1$. Its mass reads

$$m^2=\frac{26-D}{24\alpha'}. \tag{9.60}$$

Since there are $D-2$ ways to obtain $N=1$, this state is $D-2$-times degenerate. To be still Lorentz invariant, these transverse modes must form a representation of SO$(D-2)$ for a massless particle and SO$(D-1)$ for a massive particle. The former follows because there is no rest-frame for a massless particle, and the minimum momentum is at least $P^\mu=\left(E,E,\vec{0}\right)$, thus having less symmetry then the one for massive particle in the rest frame being $P^\mu=(m,\vec{0})$.

As a consequence, in $D=4$ massive bosonic particles have integer spin $j>0$ as representations of SO(3) with $2j+1$-fold degeneracies. Massless particles, however, are denoted by their helicity forming a representation of the group SO(2), having only one state with positive helicity. Because of CPT symmetry the number of states is actually doubled, since a state with positive helicity can be transformed by CPT into one with negative helicity. Put it in another way, the lowest non-trivial representation of SO(3) is 3-dimensional, a spin-1 state with three magnetic quantum numbers. For SO(2), the lowest non-trivial representation has actually only one possible magnetic quantum number, either 1 or -1. However, CPT guarantees that if one exists, then so does the other.

Going back to D dimensions there are thus $D-1$ states for massive bosonic particles, but only $D-2$ for massless ones. Since the degeneracy for the $N=1$ states is $D-2$, this implies that their mass must be zero. From this immediately follows that the theory is only Lorentz-invariant in $D=26$ dimensions, since otherwise (9.60) would not yield zero, implying also $A=-1$ due to (9.58).

Hence, this indirect inference yields that the consistency of the string theory with Lorentz and CPT invariance requires a certain number of dimensions, different to quantum field theories, which at least in principle can be formulated in any number of space-time dimensions. Note that this is actually a quantum effect, since only quantization yields the mass-dimension relation (9.60). Of course, more formal arguments can be given.

9.5.3.6 Closed string spectrum

A closed string is obtained when instead of open boundaries periodic boundaries are imposed. In this case the light-cone gauge conditions become.

$$\begin{aligned} X^\mu(\tau,L) &= X^\mu(\tau,0) \\ \partial_\sigma X^\mu(\tau,L) &= \partial_\sigma X^\mu(\tau,0) \\ \gamma_{ab}(\tau,L) &= \gamma_{ab}(\tau,0) \end{aligned}$$

Similarly, it is then possible to quantize the closed string as the open string. However, this provides another ambiguity, since the zero position of σ can now be anywhere along the string. Consequently, a shift of the zero point is another symmetry of the system as

$$\sigma' = \sigma + s(\tau).$$

To fix it requires another gauge condition, which is conveniently chosen as

$$\gamma_{\tau\sigma}(\tau,0) = 0$$

This implies that lines of constant τ are orthogonal to lines of constant σ at $\sigma=0$. This reduces the problem to translations about one string length as

$$\sigma' = \sigma + s(\tau) \mod L. \tag{9.61}$$

Nonetheless, this is sufficient to start.

Up to the formulation of the Hamiltonian then everything is as for the open string case. Of course, the solutions to the equations of motion are now different, respecting the new boundary conditions. They read

$$X^i(\tau,\sigma) = X^i + \frac{P^i}{P^+}\tau + i\left(\frac{\alpha'}{2}\right)^{\frac{1}{2}} \sum_{n=-\infty n\neq 0}^{\infty}\left(\frac{\alpha_n^i}{n}e^{-\frac{2\pi i n(\sigma+c\tau)}{L}} + \frac{\beta_n^i}{n}e^{\frac{2\pi i n(\sigma-c\tau)}{L}}\right),$$

in analogy to point quantum mechanics of a particle in a periodic box. As a consequence, there are now two independent sets of Fourier coefficients, α and β. These corresponds to oppositely directed waves along the string with α being those running in the left direction and β in the right direction.

Nonetheless, quantization proceeds as usual with the canonical quantization conditions

$$\begin{aligned} \left[Z^-, P^-\right] &= -i \\ \left[X^i, P^i\right] &= i\delta^{ij} \\ \left[\alpha_m^i, \alpha_n^j\right] &= m\delta^{ij}\delta_{m,-n} \\ \left[\beta_m^i, \beta_n^j\right] &= m\delta^{ij}\delta_{m,-n}. \end{aligned}$$

Thus, the system is again that of a set of free oscillators with a superimposed center-of-mass motion. The eigenstates are thus

$$|N,R,k\rangle = \left(\prod_{i=2}^{D-1}\prod_{n=1}^{\infty}\frac{\left(\alpha_{-n}^i\right)^{N_{in}}\left(\beta_{-n}^i\right)^{R_{in}}}{\left(n^{N_{in}}N_{in}!r^{R_{in}}R_{in}!\right)^{\frac{1}{2}}}\right)|0,0,k\rangle\,.$$

Herein N counts the number of left-moving states and R the number of right-moving states. It is then possible to obtain again the Hamiltonian in number-operator form, and to obtain the mass-shell equation as

$$m^2 = 2P^+H - P^iP^i = \frac{2}{\alpha'}\left(N+R+A+B\right),$$

and in the same way as previously also

$$A = B = \frac{2-D}{24}$$

is obtained.

However, in this case the values of N and R are restricted, since all physical states have to be invariant under the residual gauge freedom (9.61). To see this, the operator for translations on the string is useful. Essentially the gauge condition (9.61) tells that all states have to be invariant under translations along the σ direction of the string. To obtain it, the simplest starting point is the energy-momentum tensor on the world-sheet. It is given by

$$
\begin{aligned}
T^{ab} &= -4\pi\left(-\gamma\right)^{-\frac{1}{2}}\frac{\delta L}{\delta\gamma_{ab}}\\
&= -\frac{4\pi}{\left(-\gamma\right)^{\frac{1}{2}}}\frac{\delta}{\delta\gamma_{ab}}\left(-\frac{1}{4\pi\alpha'}\left(-\gamma\right)^{\frac{1}{2}}\gamma^{cd}\partial_c X^\mu\partial_d X_\mu\right)\\
&= \frac{1}{\alpha'\left(-\gamma\right)^{\frac{1}{2}}}\left(\frac{\delta\left(-\gamma\right)^{\frac{1}{2}}}{\delta\gamma_{ab}}\gamma^{cd}\partial_c X^\mu\partial_d X_\mu+\left(-\gamma\right)^{\frac{1}{2}}\frac{\delta\gamma^{cd}}{\delta\gamma_{ab}}\partial_c X^\mu\partial_d X_\mu\right)\\
&= \frac{1}{\alpha'\left(-\gamma\right)^{\frac{1}{2}}}\left(-\frac{1}{2\left(-\gamma\right)^{\frac{1}{2}}}\frac{\delta\gamma}{\delta\gamma_{ab}}\gamma^{cd}\partial_c X^\mu\partial_d X_\mu+\left(-\gamma\right)^{\frac{1}{2}}\partial_a X^\mu\partial_b X_\mu\right)\\
&= \frac{1}{\alpha'\left(-\gamma\right)^{\frac{1}{2}}}\left(-\frac{1}{2\left(-\gamma\right)^{\frac{1}{2}}}\gamma\gamma^{ab}\gamma^{cd}\partial_c X^\mu\partial_d X_\mu+\left(-\gamma\right)^{\frac{1}{2}}\partial_a X^\mu\partial_b X_\mu\right)\\
&= \frac{1}{\alpha'}\left(\partial_a X^\mu\partial_b X_\mu-\frac{1}{2}\gamma^{ab}\gamma^{cd}\partial_c X^\mu\partial_d X_\mu\right)\\
&= \frac{1}{\alpha'}\left(\partial^a X^\mu\partial^b X_\mu-\frac{1}{2}\gamma^{ab}\partial_c X^\mu\partial^c X_\mu\right).
\end{aligned}
$$

To argue that this indeed is an energy-momentum tensor[20], it is necessary to show that it has the necessary properties of an energy-momentum tensor, in particular it has to be conserved and traceless, and its $\tau\tau$-component must equal the Hamilton operator.

Start with its conservation. The elements of the energy-momentum tensors appear to be not invariant under diffeomorphisms, since the appearing expressions for γ^{ab} are not, since it seems there is no compensating factor of $\det\gamma$. However, the expression in terms of the Lagrangian is, so there must be a hidden invariance. This is in fact only possible, if the energy-momentum tensor is a constant, which would imply its conservation.

Using (9.40), this can be obtained explicitly

$$
\begin{aligned}
\partial_a T^{ab} &= \frac{1}{\alpha'}\partial_a\left(\partial^a X^\mu\partial^b X_\mu-\frac{1}{2}\gamma^{ab}\partial_c X^\mu\partial^c X_\mu\right)\\
&= \frac{1}{\alpha'}\left(\partial_a h^{ab}-\partial_a\left(\frac{1}{2}\gamma^{ab}\gamma^{cd}h_{cd}\right)\right)\\
&= \frac{1}{\alpha'}\left(\partial_a h^{ab}-\partial_a h^{ab}\right)=0,
\end{aligned}
$$

and thus the energy-momentum tensor is conserved.

[20] In quantum field theory this is already a non-obvious fact, lest in string theory.

The next condition is the one of tracelessness. Calculating the trace T^a_a explicitly yields

$$\begin{aligned}\gamma^{ab}\frac{\delta L}{\delta\gamma_{ab}} &= \frac{1}{(-\gamma)^{\frac{1}{2}}}\left(\gamma^{ab}\partial^a X^\mu\partial^b X_\mu - \partial^c X^\mu\partial_c X_\mu\right)\\ &= \frac{1}{(-\gamma)^{\frac{1}{2}}}\left(\partial^a X^\mu\partial_a X_\mu - \partial^c X^\mu\partial_c X_\mu\right) = 0 = \gamma^{ab}\frac{T^{ab}}{(-\gamma)^{\frac{1}{2}}},\end{aligned}$$

where it has been used that $\gamma^{ab}\gamma^{ab} = 2$. Finally, this yields

$$T^a_a\frac{1}{(-\gamma)^{\frac{1}{2}}} = 0,$$

confirming that the energy-momentum tensor is indeed traceless.

Using (9.40), this could also be shown more directly as

$$\begin{aligned}T^a_a &= \frac{1}{\alpha'}\left(\partial^a X^\mu\partial_a X_\mu - \frac{1}{2}\gamma^a_a\partial^c X^\mu\partial_c X_\mu\right)\\ &= \frac{1}{\alpha'}\left(\partial^a X^\mu\partial_a X_\mu - \frac{1}{2}\gamma^a_a\gamma_{cd}\partial^c X^\mu\partial^d X_\mu\right) = \frac{1}{\alpha'}\left(\partial^a X^\mu\partial_a X_\mu - \partial^a X^\mu\partial_a X_\mu\right) = 0,\end{aligned}$$

and thus the same result.

Finally, the $\tau\tau$ component should be the Hamiltonian. To show this, it is simpler to go backwards. By reexpressing the Hamiltonian (9.50) as a function of $\partial_\sigma X^\mu$ and $\partial_\tau X^\mu$ it becomes

$$\begin{aligned}H &= \frac{L}{4\pi\alpha' P^+}\int_0^L d\sigma\left(2\pi\alpha'\Pi^i\Pi^i + \frac{1}{2\pi\alpha'}\partial_\sigma X^i\partial_\sigma X^i\right)\\ &= \frac{L}{4\pi\alpha' P^+}\int_0^L d\sigma\left(2\pi\alpha'\frac{P^+}{L}\frac{P^+}{L}\partial_\tau X^i\partial_\tau X^i + \frac{1}{2\pi\alpha'}\partial_\sigma X^i\partial_\sigma X^i\right).\end{aligned}$$

Using now (9.49) changes this to

$$\begin{aligned}H &= \frac{1}{4\pi\alpha'}\int_0^L d\sigma\left(2\pi\alpha'\frac{P^+}{L}\partial_\tau X^i\partial_\tau X^i + \frac{1}{2\pi\alpha'}\frac{L}{P^+}\partial_\sigma X^i\partial_\sigma X^i\right)\\ &= \frac{1}{4\pi\alpha'}\int_0^L d\sigma\left(2\pi\alpha'\frac{\gamma_{\sigma\sigma}}{2\pi\alpha'}\partial_\tau X^i\partial_\tau X^i + \frac{1}{2\pi\alpha'}\frac{2\pi\alpha'}{\gamma_{\sigma\sigma}}\partial_\sigma X^i\partial_\sigma X^i\right)\\ &= \frac{1}{4\pi\alpha'}\int_0^L d\sigma\left(\gamma_{\sigma\sigma}\partial_\tau X^\mu\partial_\tau X_\mu + \frac{1}{\gamma_{\sigma\sigma}}\partial_\sigma X^\mu\partial_\sigma X_\mu\right).\end{aligned}$$

The expansion of i to μ in the last line was permitted because this is only an addition of zero in the second term and also a zero in the first term by virtue of the boundary conditions after exchange of integration and differentiation.

To bring the $\tau\tau$ component of the energy-momentum tensor into the same form it can be reexpressed as

$$\begin{aligned} T^{\tau\tau} &= \frac{1}{\alpha'}\left(\partial^\tau X^\mu \partial^\tau X_\mu - \gamma^{\tau\tau}\partial_\tau X^\mu \partial^\tau X_\mu - \gamma^{\tau\tau}\partial_\sigma X^\mu \partial^\sigma X_\mu\right) \\ &= \frac{1}{\alpha'}\Big(\partial^\tau X^\mu \partial^\tau X_\mu - \frac{1}{2}\left(\gamma^{\tau\tau}\gamma_{\tau\tau}\right)\partial_\tau X^\mu \partial_\tau X_\mu \\ &\quad -\frac{1}{2}\gamma^{\tau\tau}\gamma_{\tau\sigma}\partial_\tau X^\mu \partial_\sigma X_\mu - \frac{1}{2}\gamma^{\tau\tau}\gamma_{\tau\sigma}\partial_\sigma X^\mu \partial_\tau X_\mu - \frac{1}{2}\gamma^{\tau\tau}\gamma_{\sigma\sigma}\partial_\sigma X^\mu \partial_\sigma X_\mu\Big). \end{aligned}$$

Because of the gauge condition $\gamma^{\tau\tau}$ and $-\gamma^{\sigma\sigma}$ are related, and yielding that the square of $\gamma^{\tau\tau}$ is -1, because otherwise the gauge condition for the determinant would be violated, given that $\gamma_{\tau\sigma}$ vanishes. This yields

$$\begin{aligned} T^{\tau\tau} &= \frac{1}{\alpha'}\left(\frac{1}{2}\partial_\tau X^\mu \partial_\tau X_\mu + \frac{1}{2}\partial_\sigma X^\mu \partial_\sigma X_\mu\right) \\ &= \frac{1}{2\alpha'}\left(\gamma^{\sigma\sigma 2}\partial_\tau X^\mu \partial_\tau X_\mu + \partial_\sigma X^\mu \partial_\sigma X_\mu\right) = \frac{1}{2\alpha'}\gamma^{\sigma\sigma}\left(\gamma^{\sigma\sigma}\partial_\tau X^\mu \partial_\tau X_\mu + \frac{1}{\gamma^{\sigma\sigma}}\partial_\sigma X^\mu \partial_\sigma X_\mu\right), \end{aligned}$$

which concludes

$$H = -\frac{1}{2\pi}\int_0^L d\sigma \gamma^{\sigma\sigma} T^{\tau\tau}$$

where the factor $\gamma^{\sigma\sigma}$ is actually part of the measure to make the expression diffeomorphism invariant, and thus shows the correct relation between the Hamiltonian and the energy-momentum tensor.

Hence, it is permitted to use this expression for the energy-momentum tensor to obtain the operator of linear translation. It is given by the $\sigma\tau$ component, as in case of classical mechanics. Since $\gamma_{\tau\sigma} = 0$ this component is given by

$$T^{\sigma\tau} = \frac{1}{\alpha'}\left(\partial^\sigma X^\mu \partial^\tau X_\mu\right) = 2\pi c \Pi^i \partial_\sigma X^i,$$

since the $\pm$ components have a vanishing σ component each. Integrating yields the operator as

$$P = -\int_0^L d\sigma \Pi^i \partial_\sigma X^i = \frac{2\pi}{L}\left(\sum_{n=1}^{\infty}\left(\alpha^i_{-n}\alpha^i_n - \beta^i_{-n}\beta^i_n\right) + A - B\right) = \frac{2\pi}{L}\left(N - R\right).$$

Since the residual gauge freedom is essentially giving that the coordinates hop around the string by an integer times L, this can be restricted by enforcing

$$N = R. \tag{9.62}$$

Thus, the expectation value of translations along the string is zero, and any physical state has a localized coordinate system on the string. With other words, the number of left and right moving modes must be the same.

The lowest state is given again by

$$m^2 = \frac{2}{\alpha'}2\frac{2-D}{24} = \frac{2-D}{6\alpha'},$$

and is therefore again a tachyon. The lowest excited state is given by $|1,1,k\rangle$

$$m^2 = \frac{26-D}{6\alpha'}.$$

Since the degeneracy is again only $D-2$, it is not a representation of SO$(D-1)$, and therefore the state has to be massless, requiring once more $D = 26$. However, in contrast to the previous case, it is not constructed by a single creation operator with just one space-time index, but by two as

$$|1,1,k\rangle = \alpha^i_{-1}\beta^j_{-1}|0,0,k\rangle,$$

and therefore is a tensor state e_{ij}.

As in the case of large extra dimensions, this state can be separated as

$$e^{ij} = \frac{1}{2}\left(e^{ij} + e^{ji} - \frac{2}{D-2}\delta^{ij}e^{kk}\right) + \frac{1}{2}\left(e^{ij} - e^{ji}\right) + \frac{1}{D-2}\delta^{ij}e^{kk}.$$

The first term is traceless symmetric, the second antisymmetric and the third scalar. Furthermore, the occupation numbers N_{in} and R_{in} can vary freely as long as $N = R$ is fulfilled. Therefore, the number of states is doubled with respect to the open string spectrum at the same N.

To verify the assignment of spin, a little more formal investigation is useful. Note that it is always possible to obtain a spin algebra from creation and annihilation operators, when summing over oscillators, called the Schwinger representation. In case of the open string, the corresponding operators are given by

$$S^{ij} = -i\sum_{n=1}^{\infty}\frac{1}{n}\left(\alpha^i_{-n}\alpha^j_n - \alpha^j_{-n}\alpha^i_n\right) \tag{9.63}$$

and the ones for the closed string are completely analogous, just requiring that it is now necessary to sum over both, left-moving and right-moving modes. The two indices already indicate that these will be the corresponding n-dimensional generalization of the spin.

That (9.63) are indeed spin operators can be shown by explicitly calculating the corresponding algebra. Start by evaluating the commutator as

$$\begin{aligned}
\left[S^{ij}, S^{kl}\right] &= -\sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\frac{1}{nm}\left[\alpha^i_{-n}\alpha^j_n - \alpha^j_{-n}\alpha^i_n, \alpha^k_{-m}\alpha^l_m - \alpha^l_{-m}\alpha^k_m\right]\\
&= -\sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\frac{1}{nm}(\left[\alpha^i_{-n}\alpha^j_n, \alpha^k_{-m}\alpha^l_m - \alpha^l_{-m}\alpha^k_m\right]\\
&\quad -\left[\alpha^j_{-n}\alpha^i_n, \alpha^k_{-m}\alpha^l_m - \alpha^l_{-m}\alpha^k_m\right])\\
&= -\sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\frac{1}{nm}(\left[\alpha^i_{-n}\alpha^j_n, \alpha^k_{-m}\alpha^l_m\right] - \left[\alpha^i_{-n}\alpha^j_n, \alpha^l_{-m}\alpha^k_m\right]\\
&\quad -\left[\alpha^j_{-n}\alpha^i_n, \alpha^k_{-m}\alpha^l_m\right] + \left[\alpha^j_{-n}\alpha^i_n, \alpha^l_{-m}\alpha^k_m\right]).
\end{aligned}$$

It is simpler to evaluate each of the four terms individually. For this the relation

$$[ab, c] = a\,[b, c] + [a, c]\,b$$

for double commutators is quite useful, as well as the quantization conditions (9.55) are necessary. In the following the summation is kept implicit. This yields for the first term

$$\begin{aligned}
\left[\alpha^i_{-n}\alpha^j_n, \alpha^k_{-m}\alpha^l_m\right] &= \alpha^i_{-n}\left[\alpha^i_n, \alpha^k_{-m}\alpha^l_m\right] + \left[\alpha^i_{-n}, \alpha^k_{-m}\alpha^l_m\right]\alpha^j_n\\
&= \alpha^i_{-n}\alpha^k_{-m}\left[\alpha^j_n, \alpha^l_m\right] + \alpha^i_{-n}\left[\alpha^j_n, \alpha^k_{-m}\right]\alpha^l_m\\
+\alpha^k_{-m}\left[\alpha^i_{-n}, \alpha^l_m\right]\alpha^i_n + \left[\alpha^i_{-n}, \alpha^k_{-m}\right]\alpha^l_m\alpha^j_n &\\
&= \alpha^i_{-n}\alpha^k_{-m}n\delta^{jl}\delta_{n,-m} + \alpha^i_{-n}\alpha^l_m n\delta^{jk}\delta_{n,m}\\
&\quad -\alpha^k_{-m}\alpha^j_n n\delta^{il}\delta_{-n,-m} - \alpha^l_m\alpha^j_n n\delta^{ik}\delta_{-n,m}\\
&= n(\alpha^i_{-n}\alpha^k_n\delta^{jl} + \alpha^i_{-n}\alpha^l_n\delta^{jk} - \alpha^k_{-n}\alpha^j_n\delta^{il} - \alpha^l_{-n}\alpha^j_n\delta^{ik}) \qquad (9.64)
\end{aligned}$$

for the second term

$$\begin{aligned}
\left[\alpha^i_{-n}\alpha^j_n, \alpha^l_{-m}\alpha^k_m\right] &= \alpha^i_{-n}\left[\alpha^j_n, \alpha^l_{-m}\alpha^k_m\right] + \left[\alpha^i_{-n}, \alpha^l_{-m}\alpha^k_m\right]\alpha^j_n\\
&= \alpha^i_{-n}\alpha^l_{-m}\left[\alpha^j_n, \alpha^k_m\right] + \alpha^i_{-n}\left[\alpha^j_n, \alpha^l_{-m}\right]\alpha^k_m\\
&\quad +\alpha^l_{-m}\left[\alpha^i_{-n}, \alpha^k_m\right]\alpha^j_n + \left[\alpha^i_{-n}, \alpha^l_{-m}\right]\alpha^k_m\alpha^j_n\\
&= \alpha^i_{-n}\alpha^l_{-m}n\delta^{jk}\delta_{n,-m} + \alpha^i_{-n}\alpha^k_m n\delta^{jl}\delta_{n,m}\\
&\quad -\alpha^l_{-m}\alpha^j_n n\delta^{ik}\delta_{-n,-m} - \alpha^k_m\alpha^j_n n\delta^{il}\delta_{-n,m}\\
&= n(\alpha^i_{-n}\alpha^l_n\delta^{jk} + \alpha^i_{-n}\alpha^k_n\delta^{jl} - \alpha^l_{-n}\alpha^j_n\delta^{ik} - \alpha^k_{-n}\alpha^j_n\delta^{il}), \qquad (9.65)
\end{aligned}$$

for the third term

$$\begin{aligned}
\left[\alpha^j_{-n}\alpha^i_n,\alpha^k_{-m}\alpha^l_m\right] &= \alpha^j_{-n}\left[\alpha^i_n,\alpha^k_{-m}\alpha^l_m\right]+\left[\alpha^j_{-n},\alpha^k_{-m}\alpha^l_m\right]\alpha^i_n\\
&= \alpha^j_{-n}\alpha^k_{-m}\left[\alpha^i_n,\alpha^l_m\right]+\alpha^j_{-n}\left[\alpha^i_n,\alpha^k_{-m}\right]\alpha^l_m\\
&\quad+\alpha^k_{-m}\left[\alpha^j_{-n},\alpha^l_m\right]\alpha^i_n+\left[\alpha^j_{-n},\alpha^k_{-m}\right]\alpha^l_m\alpha^i_n\\
&= \alpha^j_{-n}\alpha^k_{-m}n\delta^{il}\delta_{n,-m}+\alpha^j_{-n}\alpha^l_m n\delta^{ik}\delta_{n,m}\\
&\quad-\alpha^k_{-m}\alpha^i_n n\delta^{jl}\delta_{-n,-m}-\alpha^l_m\alpha^i_n n\delta^{jk}\delta_{-n,m}\\
&= n(\alpha^j_{-n}\alpha^k_n\delta^{il}+\alpha^j_{-n}\alpha^l_n\delta^{ik}-\alpha^k_{-n}\alpha^i_n\delta^{jl}-\alpha^l_{-n}\alpha^i_n\delta^{jk}), \qquad (9.66)
\end{aligned}$$

and finally the fourth

$$\begin{aligned}
\left[\alpha^j_{-n}\alpha^i_n,\alpha^l_{-m}\alpha^k_m\right] &= \alpha^j_{-n}\left[\alpha^i_n,\alpha^l_{-m}\alpha^k_m\right]+\left[\alpha^j_{-n},\alpha^l_{-m}\alpha^k_m\right]\alpha^i_n\\
&= \alpha^j_{-n}\alpha^l_{-m}\left[\alpha^i_n,\alpha^k_m\right]+\alpha^j_{-n}\left[\alpha^i_n,\alpha^l_{-m}\right]\alpha^k_m\\
&\quad+\alpha^l_{-m}\left[\alpha^j_{-n},\alpha^k_m\right]\alpha^i_n+\left[\alpha^j_{-n},\alpha^l_{-m}\right]\alpha^k_m\alpha^i_n\\
&= \alpha^i_{-n}\alpha^l_{-m}n\delta^{ik}\delta_{n,-m}+\alpha^j_{-n}\alpha^k_m\delta^{ij}\delta_{n,m}\\
&\quad-\alpha^l_{-m}\alpha^i_n n\delta^{jk}\delta_{-n,-m}-\alpha^k_m\alpha^i_n n\delta^{jl}\delta_{-n,m}\\
&= n(\alpha^j_{-n}\alpha^l_n\delta^{ik}+\alpha^j_{-n}\alpha^k_n\delta^{ij}-\alpha^l_{-n}\alpha^i_n\delta^{ik}-\alpha^k_{-n}\alpha^i_n\delta^{jl}). \qquad (9.67)
\end{aligned}$$

Combining (9.64-9.67) permits to drop the summation over m. In addition, for every δ each term appears twice, reducing the total expression to

$$\begin{aligned}
\left[S^{ij},S^{kl}\right] &= -2\sum_{n=1}^{\infty}\frac{1}{n}(\alpha^i_{-n}\alpha^k_n\delta^{jl}+\alpha^i_{-n}\alpha^l_n\delta^{jk}+\alpha^j_{-n}\alpha^k_n\delta^{il}+\alpha^j_{-n}\alpha^l_n\delta^{ik}\\
&\quad-\alpha^k_{-n}\alpha^j_n\delta^{il}-\alpha^l_{-n}\alpha^j_n\delta^{ik}-\alpha^k_{-n}\alpha^i_n\delta^{jl}-\alpha^l_{-n}\alpha^i_n\delta^{jk}).
\end{aligned}$$

Reordering, expanding -1 to i^2, and combining terms with the same δ permits to reconstruct the spin operators. Finally, the result becomes

$$\left[S^{ij},S^{kl}\right]=2i\left(\delta^{jl}S^{ik}+\delta^{jk}S^{il}+\delta^{il}S^{jk}+\delta^{ik}S^{il}\right).$$

Thus, indeed the operators satisfy a spin algebra. If all indices are different then the commutator vanishes. Since furthermore all diagonal elements of the S^{ij} vanish only elements with the same indices remain. For example this leaves

$$\left[S^{12},S^{23}\right]=2iS^{13}.$$

The commutator hence contains always the two unequal indices in the same order. Since the spin operator is antisymmetric by definition also the correct exchange property for the arguments of the commutator is obtained, completing the construction.

To see how the helicity emerges investigate first the 23 component of the spin operator, being the one relevant in a four-dimensional sub-space. The helicity of the lowest excitation of the open string is then given by

$$\langle 1,k|\,S^{23}\,|1,k\rangle = -i\sum_{n=1}^{\infty}\frac{1}{n}\,\langle 1,k|\left(\alpha^2_{-n}\alpha^3_n - \alpha^3_{-n}\alpha^2_n\right)|1,k\rangle = \langle 1,k|\,i\,|1,k\rangle = i.$$

Thus the value is 1. For the lowest excitation of the closed string, the value is found analogously to be two. Thus the lowest excitation of the open string is a vector particle while the one of the closed string is rather a graviton, in accordance with the previous considerations.

Comparing all results a number of interesting observations are obtained. Since vector particles always harbor a gauge symmetry, the open string already furnishes a gauge theory. Since it is non-interacting, this gauge theory has to be non-interacting as well, leaving only a U(1) gauge theory. A more detailed calculation would confirm this. Therefore, it is admissible to call the state $|1,k\rangle$ a photon.

Similarly, a spin 2 particle couples to a conserved tensor current. Since the only one available is the energy-momentum tensor, the symmetric contribution of the lowest excitation of the closed string can be interpreted as a graviton. The antisymmetric particle can be given the meaning of an axion, a hypothetical weakly-coupled dark matter candidate, as it is equivalent to a so-called 2-form gauge boson. Finally, the scalar particle is then the dilaton, as in the case of large extra dimensions.

Calculating the helicity gives already the correct result for the photon and the graviton. Indeed, for the axion and the dilaton a value of zero is obtained, as they would have also in a generic quantum field theory of these particles.

It should be noted that it can be shown that a string theory turns out to be only consistent if it at least contains the closed string, with the open string being an optional addition. Thus, the graviton is there in any string theory.

9.5.3.7 Dualities

It could be easily imagined that there are many different string theories, like there are many different field theories. However, the number of consistently quantizable string theories is very limited, and only five are known today. Furthermore, it can be shown that these string theories are dual to each other.

To get an idea of the concept of dualities, note the following. The Polyakov action (9.39) can also be viewed with a different interpretation: Promoting the world-sheet indices to space-time indices and taking the indices μ to label internal degrees of freedom, then the Polyakov action just describes D massless scalar fields X_μ (with internal symmetry group SO(D − 1,1)) in two space-time dimensions with a non-trivial metric γ, which is dynamically coupled to the fields. This is an example of a duality of two theories. This also demonstrates why two-dimensional field theories have played a pivotal role in understanding string theories. Another such relation is the AdS/CFT correspondence, which states that certain classical (super)gravity theories on a so-called anti de Sitter space, a special case of a curved space-time, are dual to (super)conformal field theories.

A more typical example for a string theory is the following. Start with the closed string, now also periodic in the eigentime. The condition (9.62) implies that any solution for the open string has the form

$$X^\mu = X_L^\mu + X_R^\mu.$$

It is convenient to write the left-moving and right-moving solutions for the following as

$$\begin{aligned} X_L^\mu &= \frac{x^\mu}{2} + \frac{L^2 p^\mu}{2}(\tau - \sigma) + \frac{iL}{2}\sum_{n\neq 0}\frac{\alpha_n^\mu}{n}e^{-2in(\tau-\sigma)} \\ X_R^\mu &= \frac{x^\mu}{2} + \frac{L^2 p^\mu}{2}(\tau + \sigma) + \frac{iL}{2}\sum_{n\neq 0}\frac{\beta_n^\mu}{n}e^{-2in(\tau+\sigma)}, \end{aligned}$$

where x^μ and p^μ are the position and momentum of the center of mass, and the α and β are the Fourier coefficients of the excited modes.

Now compactify one of the dimensions on a circle of radius R. This will only affect the zero-modes, so for the following the sum of excited states is dropped. Also, only the functions in the direction of the compactified dimension, say number 25, are affected. Assume that the string is warped W times around the compact dimensions. The left-moving and right-moving solutions for the ground-state take then the form

$$\begin{aligned} X_L^{25} &= \frac{x^{25}+c}{2} + (\alpha p^{25} + WR)(\tau + \sigma) \\ X_R^{25} &= \frac{x^{25}-c}{2} + (\alpha p^{25} - WR)(\tau - \sigma) \\ X^{25} &= x^{25} + \frac{2\alpha K}{R}\tau + 2WR\sigma. \end{aligned}$$

where c is an arbitrary constant which just turns the zero-point of the world-sheet coordinates around the compactified dimensions. Also, the center of mass momentum is then quantized as in the large extra dimension theories, and given by

$$p^{25} = \frac{K}{R}$$

where K is describing the Kaluza-Klein level, and is thus enumbered by an integer.

Now a duality transformation can be performed by mapping $W \to K$ and $R \to \alpha/R$. Then the zero-mode takes the form

$$X^{25} = x^{25} + 2WR\tau + 2\frac{2\alpha K}{R}\sigma.$$

However, this is exactly the expression which would be obtained if a string would wind K times around a compact dimension of size α/R for the Wth Kaluza-Klein mode. Since these parameters only appear in the zero mode, the remaining part of the solution is the same. Hence, these two theories have the same solutions under this mapping of the parameters, they are dual to each other. Such relations are called duality relations. In general, when the exact solutions are not known like in the present case, it is much harder to establish the duality of two theories. In particular, a duality in a classical theory could be broken by quantum effects. Therefore, most dualities so far have only been conjectured on the basis that no counter-example for them is known.

Since all known, consistent quantum string theories are dual to each other, the idea that there is a common underlying structure, the mentioned M-theory, is very appealing, though unproven.