

Particle Physics of the Early Universe

Uwe-Jens Wiese
Institute for Theoretical Physics
Bern University

September 17, 2010

Contents

1	Introduction	7
2	Basics of General Relativity	11
2.1	Basics of Riemann's Geometry	11
2.2	Curved Space-Time	18
3	Standard Big Bang Cosmology	23
3.1	The Friedmann-Lemaitre-Robertson-Walker Metric	24
3.2	Solutions of the Field Equations	28
3.3	Determination of the Parameters of our Universe	34
4	Thermodynamics of a Hot Big Bang	41
4.1	Thermodynamical Distributions	42
4.2	Entropy Conservation and Neutrino Temperature	48
4.3	The End of Radiation Dominance	51
5	Nucleosynthesis	55
5.1	The Neutron-Proton Ratio	55
5.2	Computed and Observed Abundances	57

6	The Standard Model of Particle Physics	59
6.1	Scalar Electrodynamics	60
6.2	The Electroweak Interaction	64
6.3	The Leptons	70
6.4	The Strong Interactions	71
6.5	Spontaneous Chiral Symmetry Breaking	78
7	Electroweak and QCD Phase Transitions	89
7.1	The QCD Phase Transition	90
7.2	Phase Transition in the Purge Glue Theory	92
7.3	Domain Walls and Gluonic Interfaces	93
7.4	Complete versus Incomplete Wetting	94
7.5	An Effective 3-d $\mathbb{Z}(3)$ Symmetric Φ^4 Model	96
7.6	Interfaces and Critical Exponents	97
7.7	The Electroweak Phase Transition	100
8	Grand Unified Theories	101
8.1	The minimal $SU(5)$ Model	102
8.2	The Fermion Multiplets	106
8.3	Predictions of Grand Unified Theories	109
9	Baryon Asymmetry	111
9.1	Evidence for a Baryon Asymmetry	111
9.2	Necessary Conditions for a Baryon Asymmetry	112
9.3	Baryon Number Violation in the Standard Model	114

10 Topological Excitations	117
10.1 Domain Walls	118
10.2 Cosmic Strings	119
10.3 Magnetic Monopoles	122
10.4 The Kibble Mechanism	124
11 Axions	127
11.1 The U(1)-Problem	129
11.2 A Solution of the Strong CP Problem	132
11.3 Axion Properties	134
11.4 Cosmological Implications of the Axion	135
12 Inflation	137
12.1 Deficiencies of the Standard Cosmology	139
12.2 The Idea of Inflation	142
12.3 The Dynamics of Inflation	143
12.4 Supernatural Inflation	146
13 Quantum Cosmology	149
13.1 Wheeler-DeWitt Equation in Mini-Superspace	151
13.2 The Wave Function of the Universe	152
13.3 The Initial Conditions for Inflation	153
13.4 Cosmology with Extra Dimensions	154
14 Large Scale Structure	157
14.1 The Jeans Instability	158

14.2 Four Scenarios for Structure Formation	160
14.3 Four Eras in the Late Universe	161
A Quantum Field Theory	165
A.1 From Point Mechanics to Classical Field Theory	165
A.2 Path Integral in Real Time	167
A.3 Path Integral in Euclidean Time	171
A.4 Spin Models in Classical Statistical Mechanics	172
A.5 Quantum Mechanics versus Statistical Mechanics	174
A.6 Lattice Field Theory	175
B Group Theory of S_N and $SU(n)$	181
B.1 The Permutation Group S_N	181
B.2 The Group $SU(n)$	184
C Topology of Gauge Fields	189
C.1 The Anomaly	189
C.2 The Topological Charge	190
C.3 Gauge Field Topology on a Compact Manifold	194
C.4 Cochain Reduction in $SU(2)$	197
C.5 The Instanton in $SU(2)$	203
C.6 θ -Vacua	204

Chapter 1

Introduction

Physics of the Early Universe means physics 14 billion years ago. We are talking about the first few minutes after the Big Bang. How can we speak about such early times in a meaningful way? We try to understand an “experiment” that was started a long time ago, and that we cannot repeat (at least we don’t know how). Still, the creation of “Universes” in the laboratory is a matter of scientific speculation with many interesting aspects. From a linguistic point of view the creation of other Universes is paradoxical, because the word “Universe” implies that there can only be one. Can we ever know if a Big Bang has really happened, or is it just a modern version of belief in a particular act of creation? If it were, this course would not be offered. As physicists we are using the scientific method: we observe (perform experiments) and we describe our observations mathematically. The theory leads to new predictions that we can test experimentally. As long as this interaction between observation and theory does not lead to a contradiction, the theory is not defeated. We cannot ask for more, because it is impossible to completely verify a theory by experimental observation. All we can hope for is not to defeat it. Our most successful theories today are general relativity and the standard model of particle physics. Both have passed all experimental tests performed so far very well, and they are still tested further. These theories summarize all we know about the fundamental forces: gravity, electromagnetism, as well as the weak and strong interactions. When we apply general relativity and the standard model of particle physics to the early Universe, we assume that they were valid then in exactly the same form as they are today. We make this assumption, not because it is something we have to believe in, but because it is part of our scientific method. The theory makes predictions (in this case post-dictions to be more precise) and we want to test those. How do we do that? We

go outside at night, and we look at the sky. If we are an astrophysicist, we may be lucky and use one of the big telescopes on Hawaii. With that we collect light, very old light, that itself may have traveled 10 billion years, carrying information about the Universe at a very early age. When we analyze the light spectroscopically, we find the spectral lines characteristic for specific atoms — a direct indication that several billion years ago, atomic physics was working in exactly the same way as it works today.

However, the spectral lines are red-shifted, and the red-shift is larger the deeper we look in space (and hence the further back we look in time). This observation, first made by Edwin Hubble in the twenties of the past century, has drastically changed our picture of the Universe. When interpreted in terms of the Doppler effect, Hubble’s observation means that distant galaxies are moving away from us with velocities proportional to their distance. A few hundred years ago, people might have taken this as a proof that we are at the center of the Universe. Today we are more modest, and assume that everybody in the Universe would see the same effect. Hence, we conclude that the Universe as a whole — i.e. space itself — is expanding. Using the observational fact that the Universe is expanding, general relativity allows us to calculate backwards in time. This inevitably leads to a Big Bang. From the point of view of general relativity the Big Bang is a singularity, in which space and time are “created”. At ultra-early times we cannot trust classical general relativity, because then quantum effects of gravity become important. Also we do not yet have a satisfactory description of quantum gravity, although there are many interesting ideas, for example, in the framework of string theory. The typical scale of gravity is set by Newton’s constant

$$G = 6.6720 \times 10^{-8} \text{cm}^3 \text{g}^{-1} \text{sec}^{-2}. \quad (1.0.1)$$

Besides that we only have \hbar and c as fundamental constants. Using these three, we can define the Planck energy (or Planck mass)

$$m_P = \sqrt{\frac{\hbar c^5}{G}} = 1.2211 \times 10^{19} \text{GeV}. \quad (1.0.2)$$

Correspondingly, we can define the Planck time

$$t_P = \sqrt{\frac{\hbar G}{c^5}} = 5.3904 \times 10^{-44} \text{sec}. \quad (1.0.3)$$

General relativity is robust enough to be trusted at all energies below m_P . In the early Universe this means we can rely on it for all times $t > t_P$ after the Big Bang. Everything before that is speculation based on our favorite model of quantum gravity. Of course, nobody can stop us from speculating, and we

will do that at the end of the course (provided I learn enough about quantum gravity until then). At present, we cannot speak about the moment of “creation”, $t = 0$, in a meaningful way, because our concepts of space and time (as we know them) then break down. Thus, even in modern cosmology there is room for belief how the Universe may have begun, and who may have started this exciting “experiment”.

However, let us talk about the time after t_P . At that time — before the billions of years of expansion — the part of the Universe that we can observe today had a diameter of about 0.01mm. In the mean time no energy has been lost, it only got diluted. Consequently, the energy density was enormous in the early Universe — the Big Bang left us with an extremely hot system. Energy densities that we cannot generate today even with the biggest particle accelerators, were naturally available at early enough times. Hence, these were ideal times for particle physics. Even the most massive particles in the most exotic extension of the standard model may have existed immediately after the Bang. This leads to many interesting phenomena, both within and beyond the standard model, that we will discuss in detail in this course.

All we know about particle physics today is summarized in the so-called standard model, which describes the strong and electro-weak interactions as an $SU(3)_c \otimes SU(2)_L \otimes U(1)_Y$ non-Abelian gauge theory. In this theory the players are quarks and leptons (electrons and neutrinos), gauge bosons (photons, gluons and W- and Z-bosons) and Higgs particles. The standard model has been tested at energies up to several hundred GeV. This allows us to make statements about everything after the first 10^{-12} sec after the Bang. This represents an enormous progress achieved over the last twenty years. When Steven Weinberg wrote his famous book “the first three minutes”, he could only speak about those times in a meaningful way. Of course, the standard model of particle physics contains electrodynamics (and hence atomic physics) as well as nuclear physics, and those theories already give rise to very interesting phenomena. In the early Universe electrons, positrons and photons were in thermal equilibrium. The equilibrium could be maintained as long as electrons and photons interacted frequently enough. However, during the expansion of the universe, electrons and photons became more and more diluted, and interactions occurred less and less often. At about 100000 years after the Big Bang electrons and photons were decoupled. The photons from those days still travel through the Universe, and form the 2.735K cosmic background radiation, that was first observed by Penzias and Wilson in 1965, and that is an important experimental evidence for the hot beginning of our world.

In the framework of the standard model some questions about, for example, the quantization of charge or the origin of the baryon asymmetry, remain unanswered. Therefore theorists have considered models beyond the standard model, which we can presently not directly verify in experiments, simply because particle accelerators cannot reach the relevant energy scales. In these models theoretical experience and esthetical considerations are guiding principles, but we cannot be sure if we are on the right track. One such track leads to the grand unified theories (GUT), which indeed explain charge quantization via the existence of magnetic monopoles, and which also offer an explanation of the baryon asymmetry because they contain baryon number violating processes. These theories are discussed at energy scales of about 10^{15}GeV , which corresponds to times of about 10^{-36}sec after the Big Bang. Great unified theories lead to topological excitations like monopoles, cosmic strings and domain walls. Why don't we see these objects in the Universe today?

This is just one among many questions that arise in the context of the standard Big Bang cosmology. Why did our Universe reach such an old age in comparison to t_P , which sets the scale for the natural life-time of a Universe? Why is the Universe homogeneous on the largest scales, and why is the cosmic background radiation isotropic with only very small fluctuations in its temperature? On the other hand, how can we understand the large scale structures that we observe in the Universe today? A possible answer to the age problem is offered by the anthropic principle: we can only live in a Universe that gets old enough to allow the development of intelligent life forms. This argument, however, is not entirely satisfactory from a physicist's point of view. A more scientific answer to the above questions is offered by Alan Guth's idea of the inflationary Universe — a Universe that expanded exponentially at very early times. This naturally leads to a flat, homogeneous and old Universe. At the same time, inflation eliminates all kinds of unwanted topological objects like monopoles, that indeed have not been observed.

Inflation solves many problems, but many questions still remain unanswered. For example, the cosmological constant problem is as hard as ever. In this course, we first want to concentrate on those problems that are rather well understood, and we will try to distinguish them from the more speculative results. Of course, we also want to speculate — always based on theoretical models. At the end of the course we will hopefully know a little bit more about the Universe, the Universe itself will certainly just be a little bit older.

Chapter 2

Basics of General Relativity

General relativity is a theory of gravity, a generalization of Newton's theory. For weak gravitational fields and for processes slow compared to the velocity of light, it reduces to Newton's theory of gravity. General relativity is a classical theory, which has strongly resisted quantization, just like Einstein himself had problems with quantum mechanics. General relativity is physics in curved space-time. The gravitational "field" is represented by the metric of space-time, and by related quantities like the connection, Riemann's curvature tensor, the Ricci tensor and the curvature scalar. Matter fields are the sources that curve space-time, and in turn the dynamics of matter is influenced by the curvature itself. In the early Universe we don't need much of general relativity, but we must understand its basic principles, and we must get used to Riemann's geometry. First, we will talk just about space (not yet about space-time), and indeed curved spaces play a role in the early Universe, although our space seems to be rather flat.

2.1 Basics of Riemann's Geometry

In 1854 in Göttingen (Germany) Riemann formulated the theory that is used to describe the geometry of curved spaces. Half a century later, Einstein used Riemann's mathematical framework to describe curved space-time.

Let us begin with something very simple, namely with two-dimensional spaces, and let us first discuss the simplest 2-d space, the plane \mathbb{R}^2 . We use Cartesian coordinates and characterize a point by $x = (x^1, x^2)$. The distance between two

infinitesimally close points is computed following Pythagoras

$$(ds)^2 = (dx^1)^2 + (dx^2)^2 = g_{ij}(x) dx^i dx^j. \quad (2.1.1)$$

We have just introduced a metric

$$g(x) = \begin{pmatrix} g_{11}(x) & g_{12}(x) \\ g_{21}(x) & g_{22}(x) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \quad (2.1.2)$$

The metric is independent of x , which is a special case, because \mathbb{R}^2 is flat. Also the metric is symmetric, i.e.

$$g_{ji}(x) = g_{ij}(x), \quad (2.1.3)$$

which is generally the case.

Coordinates, however, are not physical. We can choose other ones if we want to. Sometimes it is useful to work with curve-linear coordinates, for example with polar coordinates. We then have

$$x^1 = r \cos \varphi, \quad x^2 = r \sin \varphi, \quad (2.1.4)$$

such that

$$dx^1 = \cos \varphi \, dr - r \sin \varphi \, d\varphi, \quad dx^2 = \sin \varphi \, dr + r \cos \varphi \, d\varphi, \quad (2.1.5)$$

and hence

$$(ds)^2 = (dr)^2 + r^2 (d\varphi)^2. \quad (2.1.6)$$

Now we have another metric

$$g(x) = \begin{pmatrix} g_{rr}(r, \varphi) & g_{r\varphi}(r, \varphi) \\ g_{\varphi r}(r, \varphi) & g_{\varphi\varphi}(r, \varphi) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}, \quad (2.1.7)$$

which does depend on x (namely on r). The metric depends on the choice of coordinates. Of course, the space is still the same. In particular, even a flat space may have an x -dependent metric. Hence, curve-linear coordinates certainly do not imply a curved space.

Let us consider something a bit more interesting, a curved space — the sphere S^2 . We can easily visualize the sphere S^2 by embedding it in the flat space \mathbb{R}^3 . However, two-dimensional inhabitants of the surface of the sphere could calculate just as we do — embeddings are not necessary. For 3-d people like ourselves it is particularly easy, because we just write

$$x^1 = R \sin \theta \cos \varphi, \quad x^2 = R \sin \theta \sin \varphi, \quad x^3 = R \cos \theta, \quad (2.1.8)$$

and therefore

$$\begin{aligned}
 (ds)^2 &= (dx^1)^2 + (dx^2)^2 + (dx^3)^2 \\
 &= R^2(\cos \theta \cos \varphi d\theta - \sin \theta \sin \varphi d\varphi)^2 \\
 &\quad + R^2(\cos \theta \sin \varphi d\theta - \sin \theta \cos \varphi d\varphi)^2 + R^2 \sin^2 \theta d\theta^2 \\
 &= R^2(d\theta^2 + \sin^2 \theta d\varphi^2).
 \end{aligned} \tag{2.1.9}$$

The metric is thus given by

$$g(x) = R^2 \begin{pmatrix} 1 & 0 \\ 0 & \sin^2 \theta \end{pmatrix}. \tag{2.1.10}$$

Of course, we can again introduce other coordinates. The following choice may seem unmotivated, but it will be useful later. We introduce

$$\rho = \sin \theta, \quad d\rho = \cos \theta d\theta, \tag{2.1.11}$$

and obtain

$$(ds)^2 = R^2 \left(\frac{d\rho^2}{1-\rho^2} + \rho^2 d\varphi^2 \right). \tag{2.1.12}$$

Thus the new metric takes the form

$$g(x) = R^2 \begin{pmatrix} \frac{1}{1-\rho^2} & 0 \\ 0 & \rho^2 \end{pmatrix}. \tag{2.1.13}$$

The sphere has a constant positive curvature. Let us now consider a space with constant negative curvature — the hyperbolically curved surface H^2 . We can no longer embed this surface in \mathbb{R}^3 , and we may thus feel a bit uncomfortable. For the 2-d people on the surface of S^2 it does not make a difference, and from now on also we must rely just on the mathematics. A possible metric of H^2 is

$$(ds)^2 = R^2(d\theta^2 + \sinh^2 \theta d\varphi^2), \tag{2.1.14}$$

where now $\theta \in [0, \infty]$. We can embed H^2 in a 3-d Minkowski-space by writing

$$(ds)^2 = (dx^1)^2 + (dx^2)^2 - (dx^3)^2, \tag{2.1.15}$$

if we identify

$$x^1 = R \sinh \theta \cos \varphi, \quad x^2 = R \sinh \theta \sin \varphi, \quad x^3 = R \cosh \theta, \tag{2.1.16}$$

because we then have

$$\begin{aligned}
 (ds)^2 &= R^2(\cosh \theta \cos \varphi d\theta - \sinh \theta \sin \varphi d\varphi)^2 \\
 &\quad + R^2(\cosh \theta \sin \varphi d\theta - \sinh \theta \cos \varphi d\varphi)^2 - R^2 \sinh^2 \theta d\theta^2 \\
 &= R^2(d\theta^2 + \sinh^2 \theta d\varphi^2).
 \end{aligned} \tag{2.1.17}$$

We can still try to visualize H^2 , when we realize that

$$(x^1)^2 + (x^2)^2 - (x^3)^2 = R^2(\sinh^2 \theta - \cosh^2 \theta) = -R^2. \quad (2.1.18)$$

Interpreting

$$(x^3)^2 = (x^1)^2 + (x^2)^2 + R^2 \quad (2.1.19)$$

correctly is, however, delicate, because we still are in Minkowski-space, and we may hence still feel a bit uncomfortable. Let us again introduce new coordinates

$$\rho = \sinh \theta, \quad d\rho = \cosh \theta \, d\theta, \quad (2.1.20)$$

such that the metric then is

$$(ds)^2 = R^2 \left(\frac{d\rho^2}{1 + \rho^2} + \rho^2 d\varphi^2 \right). \quad (2.1.21)$$

We can now write the metrics for \mathbb{R}^2 , S^2 and H^2 in a common form

$$(ds)^2 = R^2 \left(\frac{d\rho^2}{1 - k\rho^2} + \rho^2 d\varphi^2 \right). \quad (2.1.22)$$

Here $k = 0, 1, -1$ for \mathbb{R}^2 , S^2 and H^2 , respectively. For \mathbb{R}^2 the introduction of $r = R\rho$ is a bit artificial. We have introduced a scale factor R , while ρ is dimensionless.

We now have a list of all two-dimensional spaces of constant curvature. This list is complete, at least as long as we restrict ourselves to local properties. However, we can still have various global topologies. For example, we can roll up the plane to a cylinder by identifying appropriate points. Our 2-d people don't notice the difference (the cylinder is as flat as the plane), unless they travel around their entire Universe (global topology). We can also roll up the cylinder to a torus T^2 , even though we cannot visualize this in \mathbb{R}^3 without extra curvature. The torus is finite, just like the sphere, but as flat as the plane. Similarly, there are spaces that look like H^2 locally, but are finite. Spaces that are locally like S^2 are always finite. Still, we can also change the global topology of S^2 , for example by identifying antipodal points.

All we said about two-dimensional spaces of constant curvature generalizes trivially to three-dimensional spaces of constant curvature. The general form of the metric then is

$$(ds)^2 = R^2 \left(\frac{d\rho^2}{1 - k\rho^2} + \rho^2 (d\theta^2 + \sin^2 \theta \, d\varphi^2) \right). \quad (2.1.23)$$

When we investigate the Universe on the largest scales, we find that it is rather homogeneous and isotropic (galaxy distribution, cosmic background radiation). Since the observable Universe is homogeneous and isotropic, and since we assume that we don't inhabit a special place in it, we conclude that the whole Universe is a three-dimensional space of constant curvature. Note that we do not talk about space-time yet — just about 3-d space. We have investigated various metrics, and we found that the metric depends on the coordinate system. Coordinates, however, are not physical. When we want to answer physical questions (like “what is the shortest path from A to B ?”, or “how strongly curved is space ?”) we need more of Riemann's beautiful mathematics.

What is the shortest path from A to B ? We are looking for a geodesics (and that is exactly what particles in space-time do as well). Let us consider an arbitrary (d -dimensional) space with a metric, and a curve $x(\lambda)$ connecting two points A and B, i.e. $x(0) = A$, $x(1) = B$. The length S of the curve is then given by

$$S = \int_0^1 d\lambda (g_{ij}(x) \dot{x}^i \dot{x}^j)^{1/2}, \quad \dot{x} = \frac{dx}{d\lambda}. \quad (2.1.24)$$

The integral is reparameterization invariant, i.e. defining

$$x' = \frac{dx}{d\lambda'} = \frac{dx}{d\lambda} \frac{d\lambda}{d\lambda'} = \dot{x} \frac{d\lambda}{d\lambda'}, \quad (2.1.25)$$

implies

$$S' = \int_0^1 d\lambda' (g_{ij}(x) x'^i x'^j)^{1/2} = \int_0^1 d\lambda \frac{d\lambda'}{d\lambda} (g_{ij}(x) \dot{x}^i \frac{d\lambda}{d\lambda'} \dot{x}^j \frac{d\lambda}{d\lambda'})^{1/2} = S. \quad (2.1.26)$$

We are looking for a curve of shortest length — i.e. we are trying to solve a variational problem — and we can use what we know from classical mechanics. The classical action

$$S = \int_{t_0}^{t_1} dt L(x, \dot{x}) \quad (2.1.27)$$

is minimal for the classical path that solves the Euler-Lagrange equation

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{x}} - \frac{\partial L}{\partial x} = 0. \quad (2.1.28)$$

Applying this to our geodesics problem implies

$$\frac{d}{d\lambda} \frac{1}{2} (g_{ij}(x) \dot{x}^i \dot{x}^j)^{-1/2} 2g_{kl} \dot{x}^l - \frac{1}{2} (g_{ij}(x) \dot{x}^i \dot{x}^j)^{-1/2} \partial_k g_{kl} \dot{x}^l \dot{x}^m = 0. \quad (2.1.29)$$

We now use the reparameterization invariance, and parameterize the curve by its arc-length s

$$ds = (g_{ij}(x) \dot{x}^i \dot{x}^j)^{1/2} d\lambda, \quad (2.1.30)$$

such that

$$\begin{aligned} \frac{d}{ds}(g_{kl} \frac{dx^l}{ds}) - \frac{1}{2} \partial_k g_{km} \frac{dx^l}{ds} \frac{dx^m}{ds} &= 0 \Rightarrow \\ g_{kl} \frac{d^2 x^l}{ds^2} + \partial_i g_{kl} \frac{dx^i}{ds} \frac{dx^l}{ds} - \frac{1}{2} \partial_k g_{km} \frac{dx^l}{ds} \frac{dx^m}{ds} &= 0. \end{aligned} \quad (2.1.31)$$

We introduce the inverse metric

$$g^{ij} g_{jk} = \delta_k^i, \quad (2.1.32)$$

and we use the symmetry of the metric to write

$$\frac{d^2 x^l}{ds^2} + \frac{1}{2} g^{kl} (\partial_i g_{kj} + \partial_j g_{ki} - \partial_k g_{ij}) \frac{dx^i}{ds} \frac{dx^j}{ds} = 0. \quad (2.1.33)$$

The new object deserves its own name. It is called the metric connection, or just the connection

$$\Gamma_{ij}^l = \frac{1}{2} g^{kl} (\partial_i g_{kj} + \partial_j g_{ki} - \partial_k g_{ij}). \quad (2.1.34)$$

The Γ_{ij}^l are also known as Christoffel symbols. Using the connection, the equation for the geodesics takes the form

$$\frac{d^2 x^l}{ds^2} + \Gamma_{ij}^l \frac{dx^i}{ds} \frac{dx^j}{ds} = 0. \quad (2.1.35)$$

Let us consider a trivial example — geodesics in \mathbb{R}^2 . We know that these are straight lines

$$x^l = a^l s + b^l \Rightarrow \frac{d^2 x^l}{ds^2} = 0 \Rightarrow \Gamma_{ij}^l = 0. \quad (2.1.36)$$

All coefficients of the connection vanish. Now we go to polar coordinates

$$\begin{aligned} x^1 &= r \cos \varphi, \quad x^2 = r \sin \varphi \Rightarrow \\ \frac{d^2 r}{ds^2} \cos \varphi - 2 \frac{dr}{ds} \sin \varphi \frac{d\varphi}{ds} - r \cos \varphi \left(\frac{d\varphi}{ds} \right)^2 - r \sin \varphi \frac{d^2 \varphi}{ds^2} &= 0, \\ \frac{d^2 r}{ds^2} \sin \varphi + 2 \frac{dr}{ds} \cos \varphi \frac{d\varphi}{ds} - r \sin \varphi \left(\frac{d\varphi}{ds} \right)^2 + r \cos \varphi \frac{d^2 \varphi}{ds^2} &= 0 \Rightarrow \\ \frac{d^2 r}{ds^2} - r \left(\frac{d\varphi}{ds} \right)^2 &= 0, \quad 2 \frac{dr}{ds} \frac{d\varphi}{ds} + r \left(\frac{d\varphi}{ds} \right)^2 = 0 \Rightarrow \\ \Gamma_{\varphi\varphi}^r &= -r, \quad \Gamma_{r\varphi}^\varphi = \frac{1}{r}. \end{aligned} \quad (2.1.37)$$

All other connection coefficients vanish. As expected, the connection depends on the choice of coordinates. Hence, it is useful to look for a “good” coordinate

system, before one begins to compute the connection coefficients. In our example, the calculation in Cartesian coordinates was much simpler.

The connection is also used to define parallel transport of a vector along some curve. The equation

$$\dot{v}^l(\lambda) + \Gamma_{ij}^l v^i(\lambda) \dot{x}^j = 0 \quad (2.1.38)$$

yields the vector $v^l(\lambda)$ parallel transported along the curve $x(\lambda)$ for a given initial $v^l(0)$. Next we introduce a covariant derivative matrix

$$D_{ij}^l = \delta_i^l \partial_j + \Gamma_{ij}^l, \quad (2.1.39)$$

and we write

$$\begin{aligned} \dot{x}^j D_j v(x(\lambda)) &= \dot{x}^j (\delta_i^l \partial_j + \Gamma_{ij}^l) v^i(x(\lambda)) \\ &= \dot{x}^j \partial_j v^l(x(\lambda)) + \Gamma_{ij}^l v^i(x(\lambda)) \dot{x}^j \\ &= \dot{v}^l(\lambda) + \Gamma_{ij}^l v^i(\lambda) \dot{x}^j = 0. \end{aligned} \quad (2.1.40)$$

In particular, we can view the equation for a geodesics as a parallel transport equation for the tangent unit vector $t^i(s) = dx^i/ds$

$$\frac{dt^l(s)}{ds} + \Gamma_{ij}^l t^i(s) \frac{dx^j}{ds} = 0. \quad (2.1.41)$$

Parallel transport in curved space is qualitatively different from the one in flat space. Let us consider any closed curve in \mathbb{R}^2 . When we choose Cartesian coordinates, all connection coefficients vanish, and a vector parallel transported along the curve returns to its initial orientation. On the other hand, when we do the same on the surface of S^2 , the vector does not return to its initial orientation.

Let us consider parallel transport around an infinitesimal surface element. We compare the results along two paths

$$\begin{aligned} [D_j, D_k] &= D_j D_k - D_k D_j \\ &= D_{mj}^l D_{ik}^m - D_{mk}^l D_{ij}^m \\ &= (\delta_m^l \partial_j + \Gamma_{mj}^l)(\delta_i^m \partial_k + \Gamma_{ik}^m) - (\delta_m^l \partial_k + \Gamma_{mk}^l)(\delta_i^m \partial_j + \Gamma_{ij}^m) \\ &= \delta_i^l \partial_j \partial_k + \partial_j \Gamma_{ik}^l + \Gamma_{ik}^l \partial_j + \Gamma_{ij}^l \partial_k + \Gamma_{mj}^l \Gamma_{ik}^m \\ &\quad - \delta_i^l \partial_k \partial_j - \partial_k \Gamma_{ij}^l - \Gamma_{ij}^l \partial_k - \Gamma_{ik}^l \partial_j - \Gamma_{mk}^l \Gamma_{ij}^m \\ &= \partial_j \Gamma_{ik}^l - \partial_k \Gamma_{ij}^l + \Gamma_{mj}^l \Gamma_{ik}^m - \Gamma_{mk}^l \Gamma_{ij}^m = R_{ijk}^l. \end{aligned} \quad (2.1.42)$$

We have just introduced Riemann's curvature tensor

$$R_{ijk}^l = \partial_j \Gamma_{ik}^l - \partial_k \Gamma_{ij}^l + \Gamma_{mj}^l \Gamma_{ik}^m - \Gamma_{mk}^l \Gamma_{ij}^m. \quad (2.1.43)$$

In flat space the curvature tensor vanishes. This follows immediately when we choose Cartesian coordinates, for which all connection coefficients are zero.

Via index contraction we can obtain other curvature tensors from the Riemann tensor. The Ricci tensor, for example, is given by

$$R_{ik} = R_{ijk}^j, \quad (2.1.44)$$

and the curvature scalar is defined as

$$\mathcal{R} = R_i^i = g^{ki} R_{ik}. \quad (2.1.45)$$

Finally, the Einstein tensor is given by

$$G_{ik} = R_{ik} - \frac{1}{2} g_{ik} \mathcal{R}. \quad (2.1.46)$$

We now have discussed enough Riemann geometry in order to attack general relativity. What was geometry of space so far, will then turn into the geometrodynamics of space-time.

2.2 Curved Space-Time

We have learned a something about curved spaces, and we have discussed the basics of Riemann's geometry. Let us now apply these tools to curved space-time. From special relativity we know Minkowski-space — flat space-time. When we choose Cartesian coordinates its metric takes the form

$$(ds)^2 = (dt)^2 - (dx^1)^2 - (dx^2)^2 - (dx^3)^2. \quad (2.2.1)$$

Denoting $t = x_0$ we can write

$$(ds)^2 = g_{\mu\nu} dx^\mu dx^\nu, \quad g = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \quad (2.2.2)$$

We can discuss curved space-times when we allow more general (space-time-dependent) metrics. We can then use Riemann's geometry, define the connection

$$\Gamma_{\mu\nu}^\rho = \frac{1}{2} g^{\rho\sigma} (\partial_\mu g_{\sigma\nu} + \partial_\nu g_{\sigma\mu} - \partial_\sigma g_{\mu\nu}), \quad (2.2.3)$$

and write down the equation for a geodesic

$$\frac{d^2 x^\rho}{ds^2} + \Gamma_{\mu\nu}^\rho \frac{dx^\mu}{ds} \frac{dx^\nu}{ds} = 0. \quad (2.2.4)$$

Before we turn to Einstein's theory of gravity, let us look back to good old Newton's theory. There space and time are well separated, space is flat and time is absolute. Let us take a mass distribution $\rho(x)$ and let us compute its gravitational potential

$$\Phi(x) = G \int dy \frac{\rho(y)}{|x-y|}, \quad \Delta\Phi(x) = 4\pi G\rho(x). \quad (2.2.5)$$

A particle of mass m moving in this potential follows Newton's equation of motion

$$m \frac{d^2 x^l}{dt^2} = F^l = -m \frac{\partial\Phi}{\partial x^l} \Rightarrow \frac{d^2 x^l}{dt^2} + \frac{\partial\Phi}{\partial x^l} = 0. \quad (2.2.6)$$

The particle's path is a curved trajectory in flat space, with the particle moving along it as absolute time passes. Let us reinterpret this in terms of a curved Newton space-time (as it was first done by Cartan in 1923). Then absolute time would be another coordinate. We also introduce the proper time s of the particle, that it reads off from a comoving watch. Since Newton's time is absolute, we simply have $s = t$. We can therefore write Newton's equations of motion (with $t = x^0$) in the form

$$\frac{d^2 x^l}{ds^2} + \frac{\partial\Phi}{\partial x^l} \frac{dx^0}{ds} \frac{dx^0}{ds} = 0, \quad (2.2.7)$$

which we can identify as the equation for a geodesic in a curved space-time. The connection coefficients then are

$$\Gamma_{00}^l = \frac{\partial\Phi}{\partial x^l}, \quad l \in \{1, 2, 3\}, \quad (2.2.8)$$

and all other coefficients vanish. This connection cannot be derived from an underlying metric, however, it still defines a meaningful equation for geodesics in a curved Newton-Cartan space-time. We can also write down the Riemann tensor

$$R_{\mu\nu\rho}^\sigma = \partial_\nu \Gamma_{\mu\rho}^\sigma - \partial_\rho \Gamma_{\mu\nu}^\sigma + \Gamma_{\lambda\nu}^\sigma \Gamma_{\mu\rho}^\lambda - \Gamma_{\lambda\rho}^\sigma \Gamma_{\mu\nu}^\lambda. \quad (2.2.9)$$

The only non-vanishing components are

$$R_{0i0}^l = -R_{00i}^l = \partial_i \Gamma_{00}^l = \frac{\partial^2 \Phi}{\partial x^i \partial x^l}. \quad (2.2.10)$$

The Ricci tensor is then given by

$$R_{\mu\rho} = R_{\mu\nu\rho}^{\nu} \Rightarrow R_{00} = R_{0i0}^i = \frac{\partial^2 \Phi}{\partial x^{i2}} = \Delta \Phi = 4\pi G\rho, \quad (2.2.11)$$

and again all other components are zero. We cannot build the curvature scalar or the Einstein tensor in this case, because we don't have a metric. The above results can be interpreted as follows. The ordinary curved trajectories in flat space are “straight” lines (geodesics) in the curved Newton-Cartan space-time, and the mass density ρ is responsible for the curvature (Ricci tensor). From this point it is not far to go to Einstein's general relativity.

The existence of a metric is essential for Einstein's theory. The equation of motion for a massive particle that feels gravity only (free falling particle) is the equation for a geodesic

$$\frac{d^2 x^\rho}{ds^2} + \Gamma_{\mu\nu}^\rho \frac{dx^\mu}{ds} \frac{dx^\nu}{ds} = 0. \quad (2.2.12)$$

Here we have parameterized the particle's trajectory by its arc-length — the proper time s . Introducing the four-velocity

$$u^\mu = \frac{dx^\mu}{ds} \quad (2.2.13)$$

we can write

$$\frac{du^\rho}{ds} + \Gamma_{\mu\nu}^\rho u^\mu u^\nu = 0. \quad (2.2.14)$$

The four-velocity is a tangent unit-vector

$$u^2 = u^\mu u_\mu = g_{\mu\nu} u^\mu u^\nu = g_{\mu\nu} \frac{dx^\mu}{ds} \frac{dx^\nu}{ds} = \frac{(ds)^2}{(ds)^2} = 1. \quad (2.2.15)$$

The equation of motion for massless particles (photons, light rays) is simply

$$(ds)^2 = 0, \quad (2.2.16)$$

which is also known as a null-geodesics.

We still need an equation that determines the curvature of space-time based on the mass distribution. This is Einstein's field equation

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} \mathcal{R} = 8\pi G T_{\mu\nu}. \quad (2.2.17)$$

Here $T_{\mu\nu}$ is the energy-momentum tensor of matter (particles and fields). Hence, it is not just mass, but any form of energy and momentum, that determines the curvature of space-time. In a weak hour Einstein has modified his equation to

$$G_{\mu\nu} = 8\pi G T_{\mu\nu} + g_{\mu\nu} \Lambda. \quad (2.2.18)$$

Later, he considered the introduction of the cosmological constant Λ his biggest mistake. Since that time we have the problem to understand why the cosmological constant is so incredibly small.

Chapter 3

Standard Big Bang Cosmology

We now have the tools that we need to deal with curved space-time, and we have the dynamical equations for the motion of massive particles and of light rays. Also we have Einstein's field equations, which determine the space-time curvature from the energy-momentum distribution of particles and fields. Now we want to apply general relativity to the Universe as a whole. We are interested only in the largest scales, not in local structures like galaxy-clusters, galaxies, or even single stars, or planetary systems. Of course, all these objects curve space-time locally (in case of black holes even very strongly). However, we will average over all the local structures, and we assume that the whole Universe is filled homogeneously and isotropically by a gas of matter, whose "molecules" are, for example, the galaxies. At first sight, this idealization may seem very drastic. On the other hand, we know from hydrodynamics that a continuum description of gases works very well, although they have a very discontinuous structure at molecular scales. At the end it is a question for observations, how homogeneous and isotropic our Universe really is. The observed galaxy distribution is indeed rather homogeneous and isotropic when one averages out the structures of galaxy-clusters and individual galaxies. Also the cosmic background radiation is extremely isotropic, which indicates that space was homogeneous and isotropic immediately after the Big Bang.

The standard cosmological model assumes that the Universe is a homogeneous and isotropic three-dimensional space of constant curvature, however, with a time-dependent scale parameter R . The corresponding metric is the so-called Friedmann-Lemaitre-Robertson-Walker (FLRW) metric. The Einstein field equations then determine the dynamics of the Universe (expansion), when one assumes a certain energy-momentum tensor of matter. Before we study the dynamics of

space-time in detail, we will investigate the kinematics of the FLRW metric, i.e. we will play free falling observer and we will investigate the horizon.

3.1 The Friedmann-Lemaitre-Robertson-Walker Metric

Let us use some observational facts to obtain a useful ansatz for the metric of our Universe on the largest scales. Spatial homogeneity and isotropy lead us to the spaces of constant curvature. An experiment that we repeat every day tells us that we are interested in three-dimensional spaces. Thus we have the candidates \mathbb{R}^3 , S^3 and H^3 eventually endowed with some non-trivial global topology. The spatial part of the metric is then given by

$$R^2\left(\frac{d\rho^2}{1-k\rho^2} + \rho^2(d\theta^2 + \sin^2\theta d\varphi^2)\right), \quad k = 0, \pm 1, \quad (3.1.1)$$

where R is a scale parameter (of dimension length). We now use Hubble's observation that our Universe is expanding, and thus allow R to be time-dependent. Altogether, we make the FLRW ansatz

$$ds^2 = dt^2 - R(t)^2\left(\frac{d\rho^2}{1-k\rho^2} + \rho^2(d\theta^2 + \sin^2\theta d\varphi^2)\right). \quad (3.1.2)$$

We have used a special space-time coordinate system, which clearly manifests the symmetries of our ansatz. Indeed, we are in the comoving coordinate system of the "galaxy-gas", and only in this system the Universe appears homogeneous and isotropic.

We know how to compute the connection coefficients, the Riemann tensor, the Ricci tensor, the curvature scalar and the Einstein tensor. It is quite tedious to compute these quantities for the three-dimensional spaces of constant curvature. Since we have done this exercise in two-dimensions (in a homework), we simply

quote the results

$$\begin{aligned}
\Gamma_{ij}^l &= \frac{1}{2}g^{lk}(\partial_j g_{ki} + \partial_i g_{kj} - \partial_k g_{ij}), \\
\Gamma_{ij}^0 &= -\frac{\dot{R}}{R}g_{ij}, \quad \Gamma_{0j}^i = \frac{\dot{R}}{R}\delta_j^i, \\
R_{ij} &= -\left(\frac{\ddot{R}}{R} + 2\frac{\dot{R}^2}{R^2} + 2\frac{k}{R^2}\right)g_{ij}, \quad R_{00} = -3\frac{\ddot{R}}{R}, \\
\mathcal{R} &= -6\left(\frac{\ddot{R}}{R} + \frac{\dot{R}^2}{R^2} + \frac{k}{R^2}\right), \\
G_{ij} &= \left(2\frac{\ddot{R}}{R} + \frac{\dot{R}^2}{R^2} + \frac{k}{R^2}\right)g_{ij}, \quad G_{00} = 3\left(\frac{\dot{R}^2}{R^2} + \frac{k}{R^2}\right).
\end{aligned} \tag{3.1.3}$$

All other components vanish.

Let us now follow a light ray (photon) through the Universe. We ignore all interactions with matter (interstellar gas, plasma) and consider only gravitational effects. Hence, we are looking for a null-geodesics

$$ds^2 = 0. \tag{3.1.4}$$

Since the Universe is homogeneous and isotropic, we can limit ourselves to motion along the ρ -direction, and put $\theta = \varphi = 0$. We then have

$$ds^2 = dt^2 - R(t)^2 \frac{d\rho^2}{1 - k\rho^2} = 0. \Rightarrow \frac{dt}{R(t)} = \pm \frac{d\rho}{\sqrt{1 - k\rho^2}}. \tag{3.1.5}$$

We consider a light signal, that has been emitted at time $t = 0$ (say in the moment of the Big Bang) at the space-point with dimensionless coordinate $\rho = \rho_H$. We observe the signal today (at time t) at the point $\rho = 0$, such that

$$\int_0^t \frac{dt'}{R(t')} = \int_0^{\rho_H} \frac{d\rho}{\sqrt{1 - k\rho^2}}. \tag{3.1.6}$$

Today, what is the distance to the point, at which the light signal was emitted originally? Even though the Universe has expanded in the mean time ($R(t)$ has increased) the dimensionless coordinate of that point still is ρ_H . The distance to that point is given by

$$\begin{aligned}
d_H(t) &= \int_0^1 d\lambda \, (-g_{ij}\dot{x}^i\dot{x}^j)^{1/2} = \int_0^{\rho_H} d\rho \, \sqrt{-g_{\rho\rho}} = \int_0^{\rho_H} d\rho \, \frac{R(t)}{\sqrt{1 - k\rho^2}} \\
&= R(t) \int_0^t \frac{dt'}{R(t')}.
\end{aligned} \tag{3.1.7}$$

The distance to the horizon is determined by the behavior of $R(t)$ close to the Big Bang $t = 0$. Let us assume that $R(t) \propto t^n$. Then, for $n < 1$,

$$d_H(t) \propto t^n \int_0^t \frac{dt'}{t'^n} = \left[\frac{t^n}{(1-n)t^{n-1}} \right]_0^t = \frac{t}{1-n}, \quad (3.1.8)$$

i.e. the distance is finite. For $n \geq 1$, on the other hand, the distance to the horizon is infinite. In a radiation dominated Universe, as it existed when the cosmic background radiation was generated, one has $n = 1/2$. Now the Universe is dominated by nonrelativistic matter and we have $n = 2/3$. In both cases, the distance to the horizon is indeed finite.

Now let us consider the motion of a massive particle along a geodesic

$$\frac{du^\rho}{ds} + \Gamma_{\mu\nu}^\rho u^\mu u^\nu = 0, \quad (3.1.9)$$

and let us concentrate on the zero-component of this equation

$$\frac{du^0}{ds} + \Gamma_{\mu\nu}^0 u^\mu u^\nu = 0, \quad (3.1.10)$$

In the FLRW metric one has

$$\Gamma_{ij}^0 = -\frac{\dot{R}}{R} g_{ij}, \quad (3.1.11)$$

and we obtain

$$\frac{du^0}{ds} - \frac{\dot{R}}{R} g_{ij} u^i u^j = \frac{du^0}{ds} + \frac{\dot{R}}{R} |\vec{u}|^2 = 0. \quad (3.1.12)$$

The four-velocity is a tangent unit-vector

$$\begin{aligned} (u^0)^2 - |\vec{u}|^2 = 1 &\Rightarrow u^0 du^0 - |\vec{u}| d|\vec{u}| = 0 \Rightarrow \\ \frac{1}{u^0} \frac{d|\vec{u}|}{ds} + \frac{\dot{R}}{R} |\vec{u}| &= \frac{ds}{dx^0} \frac{d|\vec{u}|}{ds} + \frac{\dot{R}}{R} |\vec{u}| = \frac{\dot{u}}{u} + \frac{\dot{R}}{R} |\vec{u}| = 0 \Rightarrow \\ \frac{|\dot{u}|}{|\vec{u}|} &= -\frac{\dot{R}}{R}, \end{aligned} \quad (3.1.13)$$

such that

$$|\vec{u}| \propto R^{-1}. \quad (3.1.14)$$

The four-momentum is given by

$$p^\mu = m u^\mu, \quad (3.1.15)$$

which implies that the three-momentum $|\vec{p}| \propto R^{-1}$ is red-shifted. The ordinary three-velocity is

$$v^i = \frac{dx^i}{dt} = \frac{dx^i}{ds} \frac{ds}{dx^0} = \frac{u^i}{u^0} \Rightarrow |\vec{v}|^2 = \frac{|\vec{u}|^2}{(u^0)^2} = \frac{|\vec{u}|^2}{1 + |\vec{u}|^2} \Rightarrow |\vec{u}|^2 = \frac{|\vec{v}|^2}{1 - |\vec{v}|^2}. \quad (3.1.16)$$

Hence, for small $|\vec{v}|$ we have

$$|\vec{v}| \propto R^{-1}, \quad (3.1.17)$$

i.e. in an expanding Universe a free falling observer will ultimately come to rest in the cosmic rest frame of matter.

Let us also investigate the red-shift of light signals. We consider a signal emitted at time t_1 and coordinate ρ_1 , and observed at time t_2 at coordinate $\rho = 0$, such that

$$ds^2 = 0 \Rightarrow \frac{dt}{R(t)} = \pm \frac{d\rho}{\sqrt{1 - k\rho^2}} \Rightarrow \int_{t_1}^{t_0} \frac{dt}{R(t)} = \int_0^{\rho_1} \frac{d\rho}{\sqrt{1 - k\rho^2}}. \quad (3.1.18)$$

Let us assume that a wave-maximum was emitted at t_1 , and that the next wave-maximum was emitted at $t_1 + \delta t_1$. That maximum will then be observed at time $t_0 + \delta t_0$, such that

$$\int_{t_1 + \delta t_1}^{t_0 + \delta t_0} \frac{dt}{R(t)} = \int_0^{\rho_1} \frac{d\rho}{\sqrt{1 - k\rho^2}}. \quad (3.1.19)$$

This immediately implies

$$\int_{t_0}^{t_0 + \delta t_0} \frac{dt}{R(t)} = \int_{t_1}^{t_1 + \delta t_1} \frac{dt}{R(t)} \Rightarrow \frac{\delta t_0}{R(t_0)} = \frac{\delta t_1}{R(t_1)}. \quad (3.1.20)$$

The wave-length λ is inversely proportional to the frequency, and hence proportional to δt , such that

$$\frac{\lambda(t_0)}{R(t_0)} = \frac{\lambda(t_1)}{R(t_1)} \Rightarrow \lambda(t) \propto R(t). \quad (3.1.21)$$

The wave-lengths of light in the Universe are stretched together with space itself. In an expanding Universe one hence finds a red-shift z with

$$1 + z = \frac{\lambda(t_0)}{\lambda(t_1)} = \frac{R(t_0)}{R(t_1)}. \quad (3.1.22)$$

Let us introduce the Hubble parameter

$$H(t) = \frac{\dot{R}(t)}{R(t)}, \quad (3.1.23)$$

which measures the strength of the expansion, and let us Taylor expand about the present epoch

$$R(t) = R(t_0) + (t - t_0)\dot{R}(t_0) \Rightarrow \frac{R(t)}{R(t_0)} = 1 + (t - t_0)H(t_0), \quad (3.1.24)$$

and hence

$$1 + z = \frac{R(t_0)}{R(t_1)} = 1 - (t_1 - t_0)H(t_0) \Rightarrow z = (t_0 - t_1)H(t_0). \quad (3.1.25)$$

What is the distance to the emitting galaxy? We find

$$\begin{aligned} d &= \int_0^1 d\lambda (-g_{ij}\dot{x}^i\dot{x}^j)^{1/2} = \int_0^{\rho_1} d\rho \sqrt{-g_{\rho\rho}} = R(t_0) \int_0^{\rho_1} \frac{d\rho}{\sqrt{1 - k\rho^2}} \\ &= R(t_0) \int_{t_1}^{t_0} \frac{dt'}{R(t')} = t_0 - t_1, \end{aligned} \quad (3.1.26)$$

such that

$$z = dH(t_0). \quad (3.1.27)$$

This is Hubble's law: the observed red-shift of light emitted from a distant galaxy is proportional to the distance of the galaxy. To derive Hubble's law we have made a power series expansion about the present epoch. Consequently, there will be corrections to Hubble's law at early times. It is interesting that the leading order calculation is independent of the details of the dynamics of the Universe (radiation- or matter-domination, curved or flat, open or closed).

3.2 Solutions of the Field Equations

The Einstein field equation takes the form

$$G_{\mu\nu} = 8\pi GT_{\mu\nu}. \quad (3.2.1)$$

In the FLRW metric, the non-vanishing components of the Einstein tensor are

$$G_{ii} = (2\frac{\ddot{R}}{R} + \frac{\dot{R}^2}{R^2} + \frac{k}{R^2})g_{ii}, \quad G_{00} = 3(\frac{\dot{R}^2}{R^2} + \frac{k}{R^2}). \quad (3.2.2)$$

Consequently, the energy-momentum tensor of matter must be diagonal as well. Let us consider the energy-momentum tensor of an ideal gas

$$T_{\mu\nu} = (\rho + p)u_\mu u_\nu - pg_{\mu\nu}. \quad (3.2.3)$$

Here ρ is the density and p is the pressure of the gas, and u_μ is the four-velocity field. The tensor $T_{\mu\nu}$ is diagonal only if u_μ has a non-vanishing zeroth component only, i.e. if $u_0 = 1$. This means that, in the coordinate system of the FLRW metric, the gas of matter is at rest, and simply follows the expansion of the Universe (cosmic rest frame). Thus, we have

$$T_{ii} = -pg_{ii}, \quad T_{00} = \rho. \quad (3.2.4)$$

Density and pressure are temperature-dependent functions that are connected via an equation of state. From the space and time components we obtain two equations, the Friedmann equation

$$3\left(\frac{\dot{R}^2}{R^2} + \frac{k}{R^2}\right) = 8\pi G\rho, \quad (3.2.5)$$

as well as

$$2\frac{\ddot{R}}{R} + \frac{\dot{R}^2}{R^2} + \frac{k}{R^2} = -8\pi Gp. \quad (3.2.6)$$

Hence, one obtains

$$\begin{aligned} & \frac{d}{dt}(8\pi G\rho R^3) + 8\pi Gp \frac{d}{dt}R^3 = \\ & \frac{d}{dt}(3R\dot{R}^2 + 3kR) - \left(2\frac{\ddot{R}}{R} + \frac{\dot{R}^2}{R^2} + \frac{k}{R^2}\right)3R^2\dot{R} = \\ & 3\dot{R}^3 + 6R\dot{R}\ddot{R} + 3k\dot{R} - 6\ddot{R}R\dot{R} - 3\dot{R}^3 - 3k\dot{R} = 0 \Rightarrow \\ & \frac{d}{dt}(\rho R^3) = -p \frac{d}{dt}R^3. \end{aligned} \quad (3.2.7)$$

This equation can be identified as the first law of thermodynamics (energy conservation)

$$dE = -pdV, \quad (3.2.8)$$

where $E = \rho R^3$ is the rest energy of a given volume R^3 . Let us form a linear combination of the two original equations

$$\begin{aligned} & 6\frac{\ddot{R}}{R} + 3\frac{\dot{R}^2}{R^2} + 3\frac{k}{R^2} - 3\left(\frac{\dot{R}^2}{R^2} + \frac{k}{R^2}\right) = -8\pi G(3p + \rho) \Rightarrow \\ & \frac{\ddot{R}}{R} = -\frac{4\pi G}{3}(3p + \rho). \end{aligned} \quad (3.2.9)$$

For the matter that we know today, one has $3p + \rho > 0$ and hence $\ddot{R} < 0$, i.e. a decelerated expansion. If $3p + \rho > 0$ was realized also in the past, there necessarily

must have been a Big Bang, not earlier than the Hubble time $H(t_0)^{-1}$. We can read the Friedmann equation as an equation for the Hubble parameter

$$H^2 + \frac{k}{R^2} = \frac{8\pi G}{3}\rho \Rightarrow \frac{k}{H^2 R^2} = \frac{8\pi G\rho}{3H^2} - 1 = \Omega - 1. \quad (3.2.10)$$

We have introduced the parameter

$$\Omega = \frac{\rho}{\rho_c}, \quad \rho_c = \frac{3H^2}{8\pi G}. \quad (3.2.11)$$

Of course, $H^2 R^2 \geq 0$, such that we obtain a relation between the critical density of matter ρ_c (that depends on time via H) and the curvature k of space. For $k = 1$ space has positive curvature (S^3) and $\rho > \rho_c$. For $k = -1$ (H^3), on the other hand, $\rho < \rho_c$. Finally, if the density is critical ($\rho = \rho_c$) — i.e. if $\Omega = 1$ — space is flat (\mathbb{R}^3).

Let us consider the Universe today — a system of non-relativistic matter clustered in galaxies, which are far away from each other. Then $p/\rho \leq 10^{-6}$, such that we can neglect the pressure and put $p = 0$ in our matter dominated Universe. Using the first law of thermodynamics

$$\rho R^3 = \frac{3}{4\pi} M, \quad (3.2.12)$$

where M is the total mass contained in a sphere of radius R , one then finds

$$\begin{aligned} \frac{\ddot{R}}{R} &= -\frac{4\pi}{3}G\rho = -\frac{MG}{R^3} \Rightarrow \ddot{R}\dot{R} = -\frac{MG}{R^2}\dot{R} \Rightarrow \\ \frac{d}{dt}\left(\frac{1}{2}\dot{R}^2\right) &= \frac{d}{dt}\left(\frac{MG}{R}\right) \Rightarrow \frac{1}{2}\dot{R}^2 = \frac{MG}{R} + C. \end{aligned} \quad (3.2.13)$$

The constant C is determined from

$$2\frac{\ddot{R}}{R} + \frac{\dot{R}^2}{R^2} + \frac{k}{R^2} = -2\frac{MG}{R^3} + \frac{2MG}{R^3} + \frac{2C}{R^2} + \frac{k}{R^2} = 0, \quad (3.2.14)$$

such that $C = -k/2$ and thus

$$\dot{R} = \sqrt{\frac{2MG}{R} - k} \Rightarrow \int_0^t dt' = \int_0^{R(t)} \frac{dR}{\sqrt{\frac{2MG}{R} - k}}. \quad (3.2.15)$$

In flat space ($k = 0$) we find

$$t = \int_0^{R(t)} dR \sqrt{\frac{R}{2MG}} = \frac{2}{3} \sqrt{\frac{R(t)^3}{2MG}} \Rightarrow R(t) = \left(\frac{9}{2}MGt^2\right)^{1/3} \propto t^{2/3}. \quad (3.2.16)$$

The same behavior follows for small R , even in the cases $k = \pm 1$.

Let us now consider a negatively curved space ($k = -1$) for large values of R

$$t = \int_0^{R(t)} dR \Rightarrow R(t) = t. \quad (3.2.17)$$

Such a Universe expands forever, faster than a flat space. For a positively curved space, on the other hand, one has

$$\dot{R}^2 = \frac{2MG}{R} - 1 = 0 \Rightarrow R_{max} = 2MG, \quad (3.2.18)$$

i.e. the scale parameter reaches a maximal value R_{max} , and then the Universe contracts ending in a Big Crunch.

At very early times the Universe was filled with a gas of hot relativistic particles — it was radiation dominated. Then the equation of state is

$$p = \frac{1}{3}\rho, \quad (3.2.19)$$

and we obtain

$$\begin{aligned} \frac{d}{dt}(\rho R^3) &= -p \frac{d}{dt}R^3 = -\frac{1}{3}\rho 3R^2 \dot{R} = -\rho R^2 \dot{R} \Rightarrow \\ \dot{\rho} R^3 + 3\rho R^2 \dot{R} + \rho R^2 \dot{R} &= \dot{\rho} R^3 + 4\rho R^2 \dot{R} = 0, \end{aligned} \quad (3.2.20)$$

such that

$$\frac{d}{dt}(\rho R^4) = \dot{\rho} R^4 + 4\rho R^3 \dot{R} = 0. \quad (3.2.21)$$

Hence, we can write

$$\rho R^4 = N, \quad (3.2.22)$$

and we obtain

$$\begin{aligned} \frac{\ddot{R}}{R} &= -\frac{4\pi G}{3}(3p + \rho) = -\frac{8\pi G}{3} \frac{N}{R^4} \Rightarrow \ddot{R} \dot{R} = -\frac{8\pi G N}{3} \frac{\dot{R}}{R^3} \Rightarrow \\ \frac{d}{dt}(\frac{1}{2} \dot{R}^2) &= \frac{d}{dt}(\frac{4\pi G N}{3R^2}) \Rightarrow \frac{1}{2} \dot{R}^2 = \frac{4\pi G N}{3R^2} + C. \end{aligned} \quad (3.2.23)$$

Again, the constant follows from

$$2\frac{\ddot{R}}{R} + \frac{\dot{R}^2}{R^2} + \frac{k}{R^2} = -\frac{16\pi G N}{3R^4} + \frac{8\pi G N}{3R^4} + \frac{2C}{R^2} + \frac{k}{R^2} = -\frac{8\pi G}{3}\rho = -\frac{8\pi G N}{3R^4}, \quad (3.2.24)$$

such that $C = -k/2$ and therefore

$$\dot{R} = \sqrt{\frac{8\pi GN}{3R^2} - k}. \quad (3.2.25)$$

The Universe is radiation dominated at early times. Hence, we can assume that R is small, such that we can neglect k and obtain

$$t = \int_0^{R(t)} dR R \sqrt{\frac{3}{8\pi GN}} = \frac{1}{2} R(t)^2 \frac{3}{8\pi GN} \Rightarrow R(t) \propto t^{1/2}. \quad (3.2.26)$$

Let us also consider an empty Universe, however, with a cosmological constant Λ . Then we have $T_{\mu\nu} = 0$ and thus

$$G_{\mu\nu} = 8\pi G T_{\mu\nu} + g_{\mu\nu} \Lambda = g_{\mu\nu} \Lambda. \quad (3.2.27)$$

For the FLRW metric we have

$$G_{ii} = (2\frac{\ddot{R}}{R} + \frac{\dot{R}^2}{R^2} + \frac{k}{R^2})g_{ii}, \quad G_{00} = 3(\frac{\dot{R}^2}{R^2} + \frac{k}{R^2}), \quad (3.2.28)$$

and therefore

$$\begin{aligned} 2\frac{\ddot{R}}{R} + \frac{\dot{R}^2}{R^2} + \frac{k}{R^2} &= \Lambda, \quad 3(\frac{\dot{R}^2}{R^2} + \frac{k}{R^2}) = \Lambda \Rightarrow \\ \frac{\ddot{R}}{R} &= \frac{\Lambda}{3} \Rightarrow \ddot{R}R = \frac{\Lambda}{3}R\dot{R} \Rightarrow \frac{1}{2}\dot{R}^2 = \frac{\Lambda}{6}R^2 + C. \end{aligned} \quad (3.2.29)$$

As before we determine C

$$2\frac{\ddot{R}}{R} + \frac{\dot{R}^2}{R^2} + \frac{k}{R^2} = \frac{2\Lambda}{3} + \frac{\Lambda}{3} + \frac{2C}{R^2} + \frac{k}{R^2} = \Lambda \Rightarrow C = -\frac{k}{2}. \quad (3.2.30)$$

This implies

$$\dot{R} = \sqrt{\frac{\Lambda}{3}R^2 - k} \Rightarrow t = \int_{R(0)}^{R(t)} \frac{dR}{\sqrt{\frac{\Lambda}{3}R^2 - k}}. \quad (3.2.31)$$

First, let us consider a flat Universe ($k = 0$), such that

$$t = \int_{R(0)}^{R(t)} \frac{dR}{\sqrt{\frac{\Lambda}{3}R}} = \sqrt{\frac{3}{\Lambda}} [\log R]_{R(0)}^{R(t)} \Rightarrow R(t) = R(0) \exp(\frac{\Lambda}{3}t). \quad (3.2.32)$$

In this case, the Hubble parameter is

$$H(t) = \frac{\dot{R}(t)}{R(t)} = \frac{\Lambda}{3}, \quad (3.2.33)$$

and thus constant in time. Exponential expansion due to vacuum energy is exactly what happens in the inflationary Universe.

How would a cosmological constant affect the expansion of the Universe today? For a matter dominated Universe with cosmological constant we write

$$\begin{aligned}
2\frac{\ddot{R}}{R} + \frac{\dot{R}^2}{R^2} + \frac{k}{R^2} &= \Lambda, \quad 3\left(\frac{\dot{R}^2}{R^2} + \frac{k}{R^2}\right) = 8\pi G\rho + \Lambda \Rightarrow \\
\frac{d}{dt}(8\pi G\rho R^3) &= 3\frac{d}{dt}(\dot{R}^2 R + kR) - \Lambda\frac{d}{dt}R^3 \\
&= 3(2\ddot{R}\dot{R}R + \dot{R}^3 + k\dot{R}) - 3\Lambda R^2\dot{R} = 0 \Rightarrow \\
\rho R^3 &= \frac{3}{4\pi}M \Rightarrow \\
\frac{\ddot{R}}{R} &= \frac{\Lambda}{3} - \frac{4\pi}{3}G\rho = \frac{\Lambda}{3} - \frac{MG}{R^3} \Rightarrow \ddot{R}\dot{R} = \frac{\Lambda}{3}R\dot{R} - \frac{GM}{R^2}\dot{R} \Rightarrow \\
\frac{1}{2}\dot{R}^2 &= \frac{\Lambda}{6}R^2 + \frac{GM}{R} + C.
\end{aligned} \tag{3.2.34}$$

As before, we determine C

$$\begin{aligned}
2\frac{\ddot{R}}{R} + \frac{\dot{R}^2}{R^2} + \frac{k}{R^2} &= \frac{2\Lambda}{3} - \frac{2GM}{R^3} + \frac{\Lambda}{3} + \frac{2GM}{R^3} + \frac{2C}{R^2} + \frac{k}{R^2} = \Lambda \Rightarrow \\
C &= -\frac{k}{2} \Rightarrow \dot{R}^2 = \frac{\Lambda}{3}R^2 + \frac{2GM}{R} - k.
\end{aligned} \tag{3.2.35}$$

For $k = 0, -1$ the Universe expands forever. At the end the $(\Lambda/3)R^2$ term dominates, and the Universe expands exponentially. For $k = 1$ the expansion comes to an end ($\dot{R} = 0$) if

$$\frac{\Lambda}{3}R^2 + \frac{2GM}{R} = 1. \tag{3.2.36}$$

We obtain a static Universe, if at the same time $\ddot{R} = 0$, such that

$$\frac{\Lambda}{3} = \frac{MG}{R^3} \Rightarrow \frac{\Lambda}{3}R^2 + \frac{2\Lambda}{3}R^2 = 1 \Rightarrow R = \frac{1}{\sqrt{\Lambda}}. \tag{3.2.37}$$

The density in such an Einstein Universe is

$$\rho = \frac{3}{4\pi} \frac{M}{R^3} = \frac{1}{4\pi G} \Lambda. \tag{3.2.38}$$

To get an overview over all possible solutions, we interpret the equation

$$\frac{1}{2}\dot{R}^2 = \frac{\Lambda}{6}R^2 + \frac{GM}{R} - \frac{k}{2}, \tag{3.2.39}$$

as the energy conservation equation for classical motion in the potential

$$V(R) = -\frac{\Lambda}{6}R^2 - \frac{GM}{R}, \quad (3.2.40)$$

with total energy $-k/2$. The static Einstein Universe exists for a critical cosmological constant $\Lambda_c = 1/(3GM)^2$. For $\Lambda > \Lambda_c$ the Universe expands forever, if it does so at present. For $\Lambda < \Lambda_c$ there are two cases. For $\dot{R} > 0$ and large R the Universe expands forever, and there was a minimal radius in the past. For $\dot{R} > 0$ and small R there is also a solution with a maximal radius and a subsequent contraction that ends in a Big Crunch.

3.3 Determination of the Parameters of our Universe

We have seen that the standard Big Bang model allows for various solutions. Thus, the question arises, which solution corresponds to our Universe. To answer that question, we must discuss some observational facts about the Universe. The central issues are the age of the Universe, its present size, the energy density compared with its critical value, and the cosmic background radiation.

Let us begin with the age of the Universe. The Universe is certainly older than the oldest objects in it. What is the age of the earth and the solar system? From the observed isotope ratios of radioactive atomic nuclei, we can determine the age of rocks (e.g. of meteorites). For example, one uses the β -decay of ^{87}Rb to ^{87}Sr , which has a half-life of $\tau = 4.99 \times 10^{10}$ years. We then have

$$^{87}\text{Sr}(t) = ^{87}\text{Sr}(0) + ^{87}\text{Rb}(0)[1 - \exp(-t/\tau)]. \quad (3.3.1)$$

It is common to normalize with respect to a stable isotope (in this case ^{86}Sr) such that

$$\frac{^{87}\text{Sr}}{^{86}\text{Sr}}(t) = \frac{^{87}\text{Sr}}{^{86}\text{Sr}}(0) + \frac{^{87}\text{Rb}}{^{86}\text{Sr}}(0)[1 - \exp(-t/\tau)]. \quad (3.3.2)$$

Let us also consider the time-dependence of ^{87}Rb

$$\frac{^{87}\text{Rb}}{^{86}\text{Sr}}(t) = \frac{^{87}\text{Rb}}{^{86}\text{Sr}}(0) \exp(-t/\tau), \quad (3.3.3)$$

such that

$$\frac{^{87}\text{Sr}}{^{86}\text{Sr}}(t) = \frac{^{87}\text{Sr}}{^{86}\text{Sr}}(0) + \frac{^{87}\text{Rb}}{^{86}\text{Sr}}(t)[\exp(-t/\tau) - 1]. \quad (3.3.4)$$

One now investigates different pieces of the same material. In these pieces $^{87}\text{Sr}/^{86}\text{Sr}(0)$ should have been the same everywhere, because chemical processes

do not distinguish between the various isotopes. On the other hand, $^{87}\text{Rb}/^{86}\text{Sr}$ may be different for different pieces of the same probe, because they may have had a different “chemical history”. One can thus use the previous equation to determine t/τ . In this way the age of the solar system has been determined as $t = 4.57(3) \times 10^9$ years. Using models for stellar evolution one can estimate the age of the oldest stars as $t = 15(4) \times 10^9$ years.

In a flat matter dominated Universe we have

$$R(t) \propto t^{2/3} \Rightarrow H(t) = \frac{\dot{R}(t)}{R(t)} = \frac{2}{3t} \Rightarrow t = \frac{2}{3}H(t)^{-1}. \quad (3.3.5)$$

Estimating the age of the Universe as $t_0 > 10^{10}$ years, we obtain

$$H(t_0) < 65 \text{ km Mpc}^{-1} \text{ sec}^{-1}, \quad 1 \text{ pc} = 3.0856 \times 10^{13} \text{ km} = 3.2615 \text{ light years}. \quad (3.3.6)$$

What is the observational value of the Hubble parameter? The Hubble parameter can be determined from the observed red-shift of distant galaxies, by a fit to Hubble’s law

$$z = H(t_0)d, \quad (3.3.7)$$

which we derived for small values of z . In that derivation we have assumed that the emitting galaxy is at rest in the cosmic rest frame, and that we are also at rest in that frame. In practice, this will certainly be only approximately true. Still, the main difficulty in applying Hubble’s law comes from the uncertainty in determining the distance d of the emitting galaxy. One uses standard candles (like super novae), which, however, are rather controversial. From these determinations one obtains

$$50 \text{ km Mpc}^{-1} \text{ sec}^{-1} < H(t_0) < 100 \text{ km Mpc}^{-1} \text{ sec}^{-1}. \quad (3.3.8)$$

The fact that this agrees reasonably well with the estimate based on the age of the Universe is considered as a big success of the Big Bang model.

Major uncertainties in the determination of the Hubble parameter result from the motion of the emitting galaxy relative to the cosmic rest frame. An indication that our own galaxy is also moving comes from the dipole asymmetry of the cosmic background radiation, that results from our motion toward the Virgo cluster. The velocity of that motion is about $v = 300(100) \text{ km sec}^{-1}$. Super novae, whose brightness maxima are theoretically well understood, can be used as standard candles, i.e. one uses them to set the distance scale based on the observed brightness. In this way one obtains $H(t_0) = 60(15) \text{ km Mpc}^{-1} \text{ sec}^{-1}$.

Due to the difficulties in determining the Hubble parameter, alternative methods also play an important role. An interesting method uses the double quasar

0957+61, which is the doubled image of a single galaxy, that is produced by a gravitational lense (an intervening second galaxy). Long-time observations of the light emission over several years have revealed a time delay of 1.55(1) years between the two light paths. Based on an appropriate model for the gravitational lense one obtains $H(t_0) = 77(5)\text{km Mpc}^{-1}\text{sec}^{-1}$. Similar observations based on a quadruple gravitational image performed by the group of Paul Schechter at MIT have led to an estimate $H(t_0) = 50(15)\text{km Mpc}^{-1}\text{sec}^{-1}$. All these data will become more accurate in the future, so that the puzzle of Hubble's constant will hopefully get resolved.

Is our Universe positively or negatively curved, or is it flat? In other words: is $\Omega = \rho/\rho_c$ bigger, smaller, or equal to 1? The critical density only depends on the Hubble parameter, $\rho_c = 3H^2/8\pi G$. All we have to do is to measure the energy density of matter (and radiation) in our Universe. A large amount of matter is condensed in galaxies. Thus, we must count galaxies and estimate their mass. Let us assume a spherical mass distribution $\rho(r)$ of stars in a galaxy, such that

$$M(R) = 4\pi \int_0^R dr \, r^2 \rho(r), \quad (3.3.9)$$

is the total mass inside a sphere of radius R , and $M(\infty)$ is the mass of the whole galaxy. The gravitational potential of such a mass distribution is

$$\begin{aligned} \Delta\Phi = 4\pi G\rho &\Rightarrow \frac{d^2\Phi}{dr^2} + \frac{2}{r} \frac{d\Phi}{dr} = 4\pi G\rho \Rightarrow \\ r^2 \frac{d^2\Phi}{dr^2} + 2r \frac{d\Phi}{dr} &= \frac{d}{dr} \left(r^2 \frac{d\Phi}{dr} \right) = 4\pi G r^2 \rho \Rightarrow GM(R) = R^2 \frac{d\Phi}{dR}. \end{aligned} \quad (3.3.10)$$

Now let us assume that the stars in the galaxy are moving in circles around the center of the galaxy. Then the centrifugal force on a star of mass m with velocity v is

$$F(R) = m \frac{v^2}{R}. \quad (3.3.11)$$

The gravitational force, on the other hand, is

$$F(R) = m \frac{d\Phi}{dR} = m \frac{GM(R)}{R^2} \Rightarrow GM(R) = v^2 R. \quad (3.3.12)$$

Using the Doppler effect, one can determine the velocity v . For typical galaxies one finds rotation curves $v(R)$ that flatten off at large radii R , i.e.

$$M(R) \propto R \Rightarrow \rho(R) \propto \frac{1}{R^2}. \quad (3.3.13)$$

In the outer halo of a galaxy there is still a lot of matter, although we do not receive light from those regions. In fact, in the halo there is 10 times more dark matter than in the shining part of a typical galaxy. The discovery of so-called MACHOs (massive compact halo objects) has been very interesting. One observes gravitational lensing of a star in another galaxy due to a MACHO in the halo of our galaxy moving through the line of sight. The strength of the gravitational lense is determined by the mass of the MACHO. Astrophysicists can observe thousands of stars in a single night. In this way, they obtain informations about the density and mass distribution of MACHOs in our galaxy. One expects about one MACHO event per 1 million of stars per night. MACHO candidates are large planets (like Jupiter), white dwarfs or neutron stars (final states of stellar evolution). Present observations imply for the shining matter

$$\Omega_{shining} \approx 0.01, \quad (3.3.14)$$

and for the dark matter

$$\Omega_{dark} \approx 0.1, \quad (3.3.15)$$

We can also use the virial theorem to estimate Ω . Let us assume that the galaxies in a galaxy cluster are in equilibrium, so that we can apply the theorem. We briefly derive the virial theorem first. Newton's equations of motion take the form

$$m_i \ddot{\vec{r}}_i = \frac{\partial U}{\partial \vec{r}_i}, \quad U = \sum_{i < j} G \frac{m_i m_j}{|\vec{r}_i - \vec{r}_j|}. \quad (3.3.16)$$

The kinetic energy is

$$\begin{aligned} K &= \sum_i \frac{1}{2} m_i \dot{\vec{r}}_i^2 = \frac{d}{dt} \left(\sum_i \frac{1}{2} m_i \dot{\vec{r}}_i \cdot \vec{r}_i \right) - \sum_i \frac{1}{2} m_i \ddot{\vec{r}}_i \cdot \vec{r}_i \\ &= \frac{d}{dt} \left(\sum_i \frac{1}{2} m_i \dot{\vec{r}}_i \cdot \vec{r}_i \right) - \sum_i \frac{1}{2} \frac{\partial U}{\partial \vec{r}_i} \cdot \vec{r}_i. \end{aligned} \quad (3.3.17)$$

Averaging over large times T implies

$$\langle K \rangle = \frac{1}{T} \int_0^T dt K(t) = \sum_i \frac{1}{2} m_i \langle \vec{v}_i^2 \rangle = - \sum_i \frac{1}{2} \left\langle \frac{\partial U}{\partial \vec{r}_i} \cdot \vec{r}_i \right\rangle. \quad (3.3.18)$$

Now we have

$$\frac{\partial U}{\partial \vec{r}_i} = - \sum_{j \neq i} G \frac{m_i m_j}{|\vec{r}_i - \vec{r}_j|^3} (\vec{r}_i - \vec{r}_j) \Rightarrow \sum_i \frac{\partial U}{\partial \vec{r}_i} \cdot \vec{r}_i = - \sum_{i < j} G \frac{m_i m_j}{|\vec{r}_i - \vec{r}_j|}. \quad (3.3.19)$$

Assuming that all galaxies in the cluster have roughly the same mass m one obtains

$$\begin{aligned} \sum_i \langle \vec{v}_i^2 \rangle &= \sum_{i < j} Gm \langle \frac{1}{|\vec{r}_i - \vec{r}_j|} \rangle \Rightarrow N \langle \vec{v}^2 \rangle = \frac{N(N-1)}{2} Gm \langle |\vec{r}|^{-1} \rangle \Rightarrow \\ G(N-1)m &= \frac{2 \langle \vec{v}^2 \rangle}{\langle |\vec{r}|^{-1} \rangle}. \end{aligned} \quad (3.3.20)$$

Measuring the average velocity via the Doppler effect, one can estimate the total mass $M = Nm$ of the galaxy cluster. Based on the virial theorem one estimates

$$\Omega_{\text{virial}} \approx 0.2(1). \quad (3.3.21)$$

Similarly, one can estimate the density of matter in the Virgo galaxy cluster from our own velocity falling into that cluster. One obtains

$$\Omega_{\text{Virgo}} \approx 0.09(4). \quad (3.3.22)$$

The nucleosynthesis of light elements, that we will discuss in detail later, yields an upper limit on the baryon density

$$\Omega_{\text{baryons}} \leq 0.18. \quad (3.3.23)$$

We should not forget that the above estimates of Ω assume that all matter is clustered. A constant, homogeneously distributed mass could have large contributions to Ω , but would still have a negligible effect on the local mass density compared to the highly concentrated matter in galaxies. Candidates for homogeneously distributed energy are vacuum energy (a cosmological constant), a small (but non-zero) neutrino mass (in the standard model of particle physics neutrinos are massless), or a yet undetected massive particle. These so-called WIMPs (weakly interacting massive particles) cannot interact strongly with ordinary matter, because otherwise we would have found them already. There are good reasons to believe that $\Omega = 1$ (inflation). In that case, we would need

$$\Omega_{\text{homogeneous}} = 0.8(1), \quad (3.3.24)$$

and most of the matter in the Universe would still be undetected.

What are the arguments in favor of $\Omega = 1$? Let us assume that all matter in the Universe has already been identified (although the dark matter is not directly visible), and that $\Omega(t_0) = 0.1$. Then, what was the value of Ω at earlier times? In a matter dominated Universe we can neglect curvature at early times and we

obtain

$$\begin{aligned} R(t) \propto t^{2/3} &\Rightarrow H(t) = \frac{\dot{R}(t)}{R(t)} = \frac{2}{3} \frac{1}{t} \Rightarrow \\ \Omega(t) - 1 &= \frac{k}{H^2 R^2} \propto \frac{t^2}{t^{4/3}} = t^{2/3} = \frac{R(t)}{R(t_0)}. \end{aligned} \quad (3.3.25)$$

Hence, at earlier times Ω must have been closer to 1 than it is now. In particular, it must be close to 1 with incredible accuracy, which requires unnatural fine-tuning of the initial conditions of the Universe. It would be more natural to assume that Ω was 1 from the very beginning. Then it always remains 1, and the Universe is flat. Indeed, the idea of the inflationary Universe, that we will discuss in detail later, naturally predicts $\Omega = 1$. Many cosmologists therefore believe that our Universe is flat, and that most of the matter in the Universe has yet to be detected. If we take into account only the matter that has already been found, we would conclude that the Universe is negatively curved and would expand forever. Of course, we cannot be sure that this is correct. The Universe may just as well be positively curved, and we may all be headed for the Big Crunch.

Finally, let us consider the cosmic background radiation — the 2.735 K black body radiation, that is interpreted as a remnant of the hot Big Bang. One observes a Planck spectrum

$$\rho d\nu = 8\pi h \left(\frac{\nu}{c}\right)^3 \frac{d\nu}{\exp(h\nu/k_B T) - 1}, \quad (3.3.26)$$

where ρ is the energy density of photons with frequency ν . In thermodynamical equilibrium entropy is conserved. As we will see later, this implies

$$T \propto R^{-1}. \quad (3.3.27)$$

Due to the red-shift we also have

$$\nu \propto R^{-1}, \quad (3.3.28)$$

such that the Planck spectrum remains one during the expansion. The cosmic background radiation has been decoupled from the rest of the matter when electrons and atomic nuclei combined to neutral atoms and the Universe became transparent. This happened at energies in the eV range ($T_d \approx 4500$ K). Let us assume that after that period the Universe has been matter dominated, such that

$$\frac{T_d}{T_0} = \frac{R(t_0)}{R(t_d)} = \left(\frac{t_0}{t_d}\right)^{2/3} \approx \frac{4500}{2.7} \Rightarrow t_d \approx 10^{-5} t_0 \approx 10^5 \text{ years}. \quad (3.3.29)$$

Consequently, the cosmic background radiation has been decoupled about 100000 years after the Big Bang.

Chapter 4

Thermodynamics of a Hot Big Bang

As we have seen, the Universe expands, and hence it was smaller at earlier times. The energy was concentrated in a small region of space, and thus the temperature was high. Processes, which can be studied today only with the biggest particle accelerators, happened naturally at that time. In particular, matter and radiation have been in thermodynamical equilibrium. To describe the relevant processes appropriately, we need thermodynamics. The early Universe was radiation dominated, i.e. $\rho R^4 = N$ remains constant. The Friedmann equation with cosmological constant then gives

$$\frac{1}{2}\dot{R}^2 = \frac{4\pi GN}{R^2} - \frac{k}{2} + \frac{\Lambda}{6}R^2. \quad (4.0.1)$$

For small R we can neglect the curvature term as well as the cosmological constant, and we obtain

$$t = \frac{1}{2}R^2\sqrt{\frac{3}{8\pi GN}} \Rightarrow \frac{32\pi}{3}GN\frac{t^2}{R^4} = \frac{32\pi}{3}G\rho t^2 = 1. \quad (4.0.2)$$

The fact that the early Universe was in thermodynamical equilibrium leads to a tremendous simplification compared to the present epoch, because the equilibrium state is characterized completely by the temperature and other thermodynamical potentials (as e.g. chemical potentials).

4.1 Thermodynamical Distributions

Let us consider a system of free relativistic particles of mass m . The energy of a particle then is

$$E = \sqrt{k^2 + m^2}. \quad (4.1.1)$$

Here the momentum \vec{k} characterizes the state $|\vec{k}\rangle$ of the particle. The particle may also have a spin \vec{S} . The component of the spin in the direction of momentum is called helicity, and is given by $S_k = -S, \dots, S$, i.e. for a particle with spin S there are $g = 2S + 1$ helicities. The spin determines the statistics of particles. Particles with integer spin are bosons, while particles with half-integer spin are fermions. A particle state can now be characterized by $|\vec{k}, S_k\rangle$.

Let us first consider bosons. An arbitrary number of bosons can occupy the same state, and the grand canonical partition function takes the form

$$Z(\beta, \mu) = \text{Tr} \exp(-\beta(H - \mu N)), \quad (4.1.2)$$

where $\beta = 1/k_B T$ is the inverse temperature, and μ is the chemical potential. H and N are the Hamiltonian and the particle number operator. The trace extends over all possible states

$$\begin{aligned} Z(\beta, \mu) &= \prod_{\vec{k}, S_k} \sum_{n(\vec{k}, S_k)=0}^{\infty} \exp(-\beta(\sum_{\vec{k}, S_k} n(\vec{k}, S_k) \sqrt{k^2 + m^2} - \mu \sum_{\vec{k}, S_k} n(\vec{k}, S_k))) \\ &= \prod_{\vec{k}, S_k} \left(\sum_{n(\vec{k}, S_k)=0}^{\infty} \exp(-\beta(\sqrt{k^2 + m^2} - \mu)n(\vec{k}, S_k)) \right) \\ &= \prod_{\vec{k}, S_k} \frac{1}{1 - \exp(-\beta(\sqrt{k^2 + m^2} - \mu))}. \end{aligned} \quad (4.1.3)$$

Let us compute the expectation value of the particle number

$$\begin{aligned}
\langle N \rangle &= \frac{1}{Z(\beta, \mu)} \text{Tr} N \exp(-\beta(H - \mu N)) \\
&= \frac{1}{\beta} \frac{\partial}{\partial \mu} \log Z(\beta, \mu) \\
&= \frac{1}{\beta} \frac{\partial}{\partial \mu} \sum_{\vec{k}, S_k} \log \frac{1}{1 - \exp(-\beta(\sqrt{k^2 + m^2} - \mu))} \\
&= \frac{1}{\beta} g \sum_{\vec{k}} \frac{1}{1 - \exp(-\beta(\sqrt{k^2 + m^2} - \mu))} \exp(-\beta(\sqrt{k^2 + m^2} - \mu)) \beta \\
&= g \sum_{\vec{k}} \frac{1}{\exp(\beta(\sqrt{k^2 + m^2} - \mu)) - 1}
\end{aligned} \tag{4.1.4}$$

Up to now we have summed over momenta, i.e. we have assumed that we are in a finite volume L^3 , e.g. with periodic boundary conditions, and thus with momenta

$$\vec{k} = \frac{2\pi}{L} \vec{n}, \quad \vec{n} \in \mathbb{Z}^3. \tag{4.1.5}$$

In the infinite volume limit we obtain

$$\sum_{\vec{k}} \rightarrow \left(\frac{L}{2\pi}\right)^3 \int d^3k, \tag{4.1.6}$$

such that the particle density takes the form

$$n = \frac{\langle N \rangle}{L^3} = \frac{g}{(2\pi)^3} \int d^3k f(\vec{k}), \tag{4.1.7}$$

where

$$f(\vec{k}) = \frac{1}{\exp(\beta(\sqrt{k^2 + m^2} - \mu)) - 1} \tag{4.1.8}$$

is the Bose-Einstein distribution. Correspondingly, one finds for energy density and pressure

$$\begin{aligned}
\rho &= \frac{\langle H \rangle}{L^3} = \frac{g}{(2\pi)^3} \int d^3k \sqrt{k^2 + m^2} f(\vec{k}), \\
p &= \frac{\langle H \rangle}{L^3} = \frac{g}{(2\pi)^3} \int d^3k \frac{\vec{k}^2}{\sqrt{k^2 + m^2}} f(\vec{k}).
\end{aligned} \tag{4.1.9}$$

Let us repeat the calculation for fermions. Since at most one fermion can occupy a given quantum state, one obtains

$$\begin{aligned} Z(\beta, \mu) &= \prod_{\vec{k}, S_k} \left(\sum_{n(\vec{k}, S_k)=0}^1 \exp(-\beta(\sqrt{k^2 + m^2} - \mu)n(\vec{k}, S_k)) \right) \\ &= \prod_{\vec{k}, S_k} (1 + \exp(-\beta(\sqrt{k^2 + m^2} - \mu))). \end{aligned} \quad (4.1.10)$$

The corresponding expectation value of the particle number is

$$\begin{aligned} \langle N \rangle &= \frac{1}{\beta} \frac{\partial}{\partial \mu} \sum_{\vec{k}, S_k} \log(1 + \exp(-\beta(\sqrt{k^2 + m^2} - \mu))) \\ &= \frac{1}{\beta} g \sum_{\vec{k}} \frac{1}{1 + \exp(-\beta(\sqrt{k^2 + m^2} - \mu))} \exp(-\beta(\sqrt{k^2 + m^2} - \mu)) \beta \\ &= g \sum_{\vec{k}} \frac{1}{\exp(\beta(\sqrt{k^2 + m^2} - \mu)) + 1} \end{aligned} \quad (4.1.11)$$

From this we obtain the particle density

$$n = \frac{\langle N \rangle}{L^3} = \frac{g}{(2\pi)^3} \int d^3k f(\vec{k}), \quad (4.1.12)$$

now with the Fermi-Dirac distribution

$$f(\vec{k}) = \frac{1}{\exp(\beta(\sqrt{k^2 + m^2} - \mu)) + 1}. \quad (4.1.13)$$

Massless particles require a slightly different treatment. They have an energy $E = |\vec{k}|$, but they only come with two helicities $S_k = \pm S$. Similarly, chiral fermions (neutrinos) only come with one handedness ($g = 1$). In a radiation dominated Universe, we find the following equation of state

$$\rho = \frac{g}{(2\pi)^3} \int d^3k k f(\vec{k}), \quad p = \frac{g}{(2\pi)^3} \int d^3k \frac{k}{3} f(\vec{k}) = \frac{\rho}{3}. \quad (4.1.14)$$

Let us also consider the non-relativistic limit of the Bose-Einstein and Fermi-Dirac distributions. Then $T \ll m$ and hence

$$f(\vec{k}) = \frac{1}{\exp(\beta(\sqrt{k^2 + m^2} - \mu)) \pm 1} \sim \exp(-\beta(m - \mu)) \exp(-\beta \frac{k^2}{2m}). \quad (4.1.15)$$

This implies

$$n = g \left(\frac{mT}{2\pi} \right)^{3/2} \exp(-\beta(m - \mu)), \quad \rho = mn, \quad p = Tn \ll \rho. \quad (4.1.16)$$

In the following, we will be most interested in the very early radiation dominated Universe, i.e. we will concentrate on high temperatures, at which all degrees of freedom become relativistic. Then we have

$$f(\vec{k}) = \frac{1}{\exp(\beta(k - \mu)) \pm 1}. \quad (4.1.17)$$

Let us begin with the photons. For them $\mu = 0$ and hence

$$\begin{aligned} n &= \frac{2}{(2\pi)^3} 4\pi \int_0^\infty dk \, k^2 \frac{1}{\exp(\beta k) - 1} = \frac{4}{(2\pi)^2} \frac{1}{\beta^3} \Gamma(3) \zeta(3) = \frac{2}{\pi^2} \zeta(3) T^3, \\ \zeta(3) &= 1.202, \\ \rho &= \frac{2}{(2\pi)^3} 4\pi \int_0^\infty dk \, k^2 \frac{k}{\exp(\beta k) - 1} = \frac{4}{(2\pi)^2} \frac{1}{\beta^4} \Gamma(4) \zeta(4) = \frac{\pi^2}{15} T^4, \\ p &= \frac{\rho}{3} = \frac{\pi^2}{45} T^4. \end{aligned} \quad (4.1.18)$$

Let us now consider fermions with a negligible chemical potential ($\mu \ll T$), and with spin 1/2 ($g = 2$ for Dirac fermions). Then

$$\begin{aligned} n &= \frac{2}{(2\pi)^3} 4\pi \int_0^\infty dk \, k^2 \frac{1}{\exp(\beta k) + 1} = \frac{4}{(2\pi)^2} \frac{1}{\beta^3} \left(1 - \frac{1}{4}\right) \Gamma(3) \zeta(3) \\ &= \frac{3}{2\pi^2} \zeta(3) T^3, \\ \rho &= \frac{2}{(2\pi)^3} 4\pi \int_0^\infty dk \, k^2 \frac{k}{\exp(\beta k) + 1} = \frac{4}{(2\pi)^2} \frac{1}{\beta^4} \left(1 - \frac{1}{8}\right) \Gamma(4) \zeta(4) \\ &= \frac{7\pi^2}{120} T^4, \\ p &= \frac{\rho}{3} = \frac{7\pi^2}{360} T^4. \end{aligned} \quad (4.1.19)$$

Up to now, we have only considered particles and no anti-particles. Those behave exactly like the particles, and hence just increase the number of degrees of freedom.

Let us get an overview of the various elementary particles, that have determined the physics of our Universe since the first few seconds. We have the photon, leptons, and hadrons. Among the leptons are charged electrons, positrons, muons, as well as neutral neutrinos. The hadrons are baryons (protons and neutrons) or mesons (pions). Table 4.1 summarizes some particle properties. There are various conserved quantities: the electric charge Q , the baryon number B , the electron-lepton number L_e , and the muon-lepton number L_μ . The values of these quantities are conserved, and hence must be prescribed by initial conditions. This can be done by specifying chemical potentials, which fix the thermal

Species	Particle	S	g	Q	B	L_e	L_μ	M [MeV]	τ
Photon	γ	1	2	0	0	0	0	$< 3 \cdot 10^{-33}$	stable
Leptons	e^-	1/2	2	-1	0	1	0	0.511	$> 2 \cdot 10^{22}$ y
	e^+	1/2	2	1	0	-1	0	0.511	$> 2 \cdot 10^{22}$ y
	ν_e	1/2	1	0	0	1	0	$< 5 \cdot 10^{-5}$	stable
	$\bar{\nu}_e$	1/2	1	0	0	-1	0	$< 5 \cdot 10^{-5}$	stable
	μ^-	1/2	2	-1	0	0	1	105.7	$2.2 \cdot 10^{-6}$ s
	μ^+	1/2	2	1	0	0	-1	105.7	$2.2 \cdot 10^{-6}$ s
	ν_μ	1/2	1	0	0	0	1	< 0.25	stable
	$\bar{\nu}_\mu$	1/2	1	0	0	0	-1	< 0.25	stable
Baryons	p	1/2	2	1	1	0	0	938.3	$> 10^{32}$ y
	\bar{p}	1/2	2	-1	-1	0	0	938.3	$> 10^{32}$ y
	n	1/2	2	0	1	0	0	939.6	898 s
	\bar{n}	1/2	2	0	-1	0	0	939.6	898 s
Mesons	π^+	0	1	1	0	0	0	139.6	$2.6 \cdot 10^{-8}$ s
	π^0	0	1	0	0	0	0	135.0	$8.7 \cdot 10^{-17}$ s
	π^-	0	1	-1	0	0	0	139.6	$2.6 \cdot 10^{-8}$ s

Table 4.1: *Summary of particle properties.*

averages of the conserved quantities. When particles interact with each other, e.g. $i + j \rightarrow k + l$, and if they are in “chemical” equilibrium, one finds for their chemical potentials

$$\mu_i + \mu_j = \mu_k + \mu_l. \quad (4.1.20)$$

As long as photons are not Bose condensed — as we can assume for the ones in the cosmic background radiation — their chemical potential vanishes, i.e. $\mu_\gamma = 0$. Since particles and anti-particles can annihilate each other into photons, their chemical potentials are equal and opposite

$$i + \bar{i} \rightarrow n \gamma \Rightarrow \mu_i + \mu_{\bar{i}} = n\mu_\gamma = 0 \Rightarrow \mu_{\bar{i}} = -\mu_i. \quad (4.1.21)$$

Let us consider some particle interactions

$$\begin{aligned}
\mu^- &\rightarrow e^- + \bar{\nu}_e + \nu_\mu \Rightarrow \mu_{\mu^-} = \mu_{e^-} - \mu_{\nu_e} + \mu_{\nu_\mu}, \\
n &\rightarrow p + e^- + \bar{\nu}_e \Rightarrow \mu_n = \mu_p + \mu_{e^-} - \mu_{\nu_e}, \\
p + \mu^- &\rightarrow n + \nu_\mu \Rightarrow \mu_{\mu^-} = \mu_n - \mu_p + \mu_{\nu_\mu} = \mu_{e^-} - \mu_{\nu_e} + \mu_{\nu_\mu}.
\end{aligned} \quad (4.1.22)$$

We can choose four independent chemical potentials, e.g. μ_p , μ_{e^-} , μ_{ν_e} and μ_{ν_μ} . We have four initial conditions for the densities n_Q , n_B , n_{L_e} and n_{L_μ} . These

densities can be compared with the photon density n_γ . Due to electric charge neutrality of the Universe we have $n_Q = 0$. Furthermore, in our Universe $n_B/n_\gamma = 10^{-9}$, such that we can put $n_B = 0$ to a good approximation. We do not know much about the lepton asymmetries

$$n_{L_e} = n_{e^-} + n_{\nu_e} - n_{e^+} - n_{\bar{\nu}_e}, \quad n_{L_\mu} = n_{\mu^-} + n_{\nu_\mu} - n_{\mu^+} - n_{\bar{\nu}_\mu}. \quad (4.1.23)$$

However, it is natural to assume that they are of the same order of magnitude as n_B , such that we can also put $n_{L_e} = n_{L_\mu} = 0$. In that case all chemical potentials vanish.

Different particle species will remain in thermal equilibrium, only if they interact with each other often enough. Since the Universe expands, particle densities become smaller and smaller, and ultimately the various particle species decouple from each other. Still, it may be useful to assign a temperature T_i to each particle species i . We compare these temperatures with the one of the photons $T = T_\gamma$, which today is 2.735 K. For the energy density we need to consider the relativistic particles only. Then

$$\rho = \frac{\pi^2}{30} g_* T^4, \quad g_* = \sum_{\text{bosons}} g_i \left(\frac{T_i}{T}\right)^4 + \sum_{\text{fermions}} \frac{7}{8} g_i \left(\frac{T_i}{T}\right)^4. \quad (4.1.24)$$

Let us compute g_* for the standard model at a temperature of 1 TeV. Then all particles are in thermal equilibrium, and we have a single temperature T and

$$g_* = \sum_{\text{bosons}} g_i + \sum_{\text{fermions}} \frac{7}{8} g_i. \quad (4.1.25)$$

We have the following degrees of freedom. Vector-bosons: photon $g_\gamma = 2$, W -bosons $g_W = 4$, Z -boson $g_Z = 2$, gluons $g_G = 16$. Scalar boson: Higgs $g_\Phi = 4$. Thus, $g_{\text{bosons}} = 28$. Then we have leptons: electron $g_e = 4$, electron-neutrino $g_{\nu_e} = 2$, muon $g_\mu = 4$, muon-neutrino $g_{\nu_\mu} = 2$, tau $g_\tau = 4$, tau-neutrino $g_{\nu_\tau} = 2$. Quarks: up-quark $g_u = 12$, down-quarks $g_d = 12$, charm-quarks $g_c = 12$, strange-quarks $g_s = 12$, top-quark $g_t = 12$, bottom-quark $g_b = 12$. Altogether we have $g_{\text{fermions}} = 90$ and thus

$$g_* = 28 + \frac{7}{8} 90 = \frac{427}{4}. \quad (4.1.26)$$

Let us discuss the decoupling of a non-relativistic particle species. For a non-relativistic particle we have

$$f(\vec{k}) = \exp(-\beta(m - \mu)) \exp(-\beta \frac{k^2}{2m}),$$

$$n = g \left(\frac{mT}{2\pi}\right)^{3/2} \exp(-\beta(m - \mu)), \quad \rho = mn, \quad p = Tn \ll \rho. \quad (4.1.27)$$

Due to particle number conservation we have

$$n \propto R^{-3}. \quad (4.1.28)$$

The momenta of non-relativistic particles are red-shifted

$$k \propto R^{-1} \Rightarrow \frac{k^2}{2m} \propto R^{-2}. \quad (4.1.29)$$

Consequently, the temperature is red-shifted as

$$T \propto R^{-2}. \quad (4.1.30)$$

To maintain $n \propto R^{-3}$ we must also have

$$m - \mu \propto R^2, \quad (4.1.31)$$

such that indeed

$$n \propto T^{3/2} \propto R^{-3}. \quad (4.1.32)$$

4.2 Entropy Conservation and Neutrino Temperature

The second law of thermodynamics states that entropy never decreases — and indeed usually increases. Only in thermodynamical equilibrium entropy remains unchanged (adiabatic processes). After inflation, one can assume an adiabatic expansion of the Universe, such that entropy is conserved. This may not be the case, if there are strong first order phase transitions. The QCD phase transition, which is the last one in the evolution of the Universe, seems to be rather weak, such that entropy conservation should be a good approximation certainly after the first milli-second after the Bang. The entropy in a comoving volume R^3 is

$$S = \frac{\rho + p}{T} R^3, \quad (4.2.1)$$

and for a relativistic gas one obtains

$$S = \frac{4}{3} \frac{R^3}{T} \rho. \quad (4.2.2)$$

We have seen that in a radiation dominated Universe $\rho R^4 = N$ is fixed during the expansion, i.e.

$$T = \frac{4}{3} \frac{N}{S} \frac{1}{R}, \quad (4.2.3)$$

and the temperature is red-shifted as R^{-1} . Let us now look at the entropy density

$$s = \frac{S}{R^3} = \frac{\rho + p}{T} = \frac{4}{3} \frac{\rho}{T}. \quad (4.2.4)$$

If the various particle species again have different temperatures we obtain

$$s = \frac{2\pi^2}{45} g_{*s} T^3, \\ g_{*s} = \sum_{bosons} g_i \left(\frac{T_i}{T}\right)^3 + \sum_{fermions} \frac{7}{8} g_i \left(\frac{T_i}{T}\right)^3. \quad (4.2.5)$$

Let us now use conservation of entropy to determine the temperature of the neutrinos in the Universe. We go back to about 1 sec after the Big Bang, to temperatures in the MeV range. Then muons and τ -leptons have annihilated with their anti-particles, and we have a system of electrons, positrons, photons and neutrinos. The neutrinos are no longer in thermal equilibrium in this moment, because they interact only weakly. A relevant process necessary to maintain thermal equilibrium is, for example, electron-positron annihilation into a neutrino-anti-neutrino pair, whose cross section is determined by the four-fermi coupling constant G_F^2 . In the standard model the above process proceeds via the Z_0 -boson channel. Thus, the order of magnitude of G_F is given by the Z_0 -mass $m_{Z_0} = 93$ GeV

$$G_F \approx \frac{1}{m_{Z_0}^2}. \quad (4.2.6)$$

The interaction rate for the above process is

$$\sigma_W \propto G_F^2 T^2, \quad (4.2.7)$$

and the particle density of the electrons is $n_e \propto T^3$. The Hubble parameter is given by

$$H = \frac{\dot{R}}{R} \propto \sqrt{G\rho} \propto \sqrt{G} T^2. \quad (4.2.8)$$

To decide if neutrinos are in thermal equilibrium we must compare the ratio of reaction their rate and the expansion rate of the Universe (the Hubble parameter)

$$\frac{\sigma_W n_e}{H} \propto \frac{G_F^2 T^5}{\sqrt{G} T^2} \propto \frac{m_P}{m_{Z_0}^4} T^3 \approx \frac{10^{19} \text{ GeV}}{10^8 \text{ GeV}^4} T^3. \quad (4.2.9)$$

This ratio is around 1 for $T \approx 1$ MeV. At lower temperatures neutrino interactions are no longer in equilibrium. Before that time neutrinos, electrons, positrons and photons have the same temperature T . When electrons and positrons annihilate at temperatures around 0.5 MeV, their entropy goes into the photons, which

thus get heated up to a higher temperature T_γ . Before the electron-positron annihilation we have

$$S = \frac{2\pi^2}{45}(g_\gamma + \frac{7}{8}(g_e + g_{\nu_e} + g_{\nu_\mu} + g_{\nu_\tau}))(TR)^3 = \frac{4\pi^2}{45}(1 + \frac{7}{8}5)(TR)^3. \quad (4.2.10)$$

This corresponds to times about when the neutrinos decouple

$$t = \sqrt{\frac{3}{32\pi G\rho}} \approx \frac{1}{\sqrt{G}} \frac{1}{T^2} = t_P \left(\frac{m_P}{T}\right)^2 = 10^{-44} \text{sec} \left(\frac{10^{19} \text{GeV}}{10^{-3} \text{GeV}}\right)^2 = 1 \text{sec}. \quad (4.2.11)$$

After the electron-positron annihilation all entropy is in photons and neutrinos. The neutrino temperature has decreased to T_ν and the scale parameter is now R' such that

$$T_\nu R' = TR. \quad (4.2.12)$$

The entropy is then given by

$$\begin{aligned} S &= \frac{2\pi^2}{45}(g_\gamma(T_\gamma R')^3 + \frac{7}{8}(g_{\nu_e} + g_{\nu_\mu} + g_{\nu_\tau})(T_\nu R')^3) \\ &= \frac{4\pi^2}{45}((T_\gamma R')^3 + \frac{7}{8}3(T_\nu R')^3). \end{aligned} \quad (4.2.13)$$

Using entropy conservation we obtain

$$\begin{aligned} (T_\gamma R')^3 + \frac{7}{8}3(T_\nu R')^3 &= (1 + \frac{7}{8}5)(T_\nu R')^3 \Rightarrow \\ \left(\frac{T_\gamma}{T_\nu}\right)^3 &= 1 + \frac{7}{8}2 = \frac{11}{4} \Rightarrow T_\nu = \left(\frac{4}{11}\right)^{1/3} T_\gamma. \end{aligned} \quad (4.2.14)$$

During the following expansion of the Universe these temperatures are simply red-shifted, but their ratio remains fixed. Of course, the photons still interact with charged matter until they decouple about 100000 years after the Bang, when neutral atoms are formed. The number of photons is, however, much larger than the number of charged particles, such that the interactions cannot change the temperature of the photons. Since one observes the cosmic background radiation at a temperature $T_\gamma = 2.7$ K, we expect a cosmic neutrino background of temperature $T_\nu = 1.9$ K. Unfortunately, neutrinos interact so weakly that the cosmic neutrino background radiation has not yet been detected. Detecting it would be very interesting, because it would tell us something about the lepton number asymmetry of the Universe.

Let us also compute the temperature of the cosmic background radiation of gravitons. Since gravity is much weaker than the weak interactions, there is little hope that the graviton background will ever be detected. The interaction rate

of gravitons is proportional to $G^2 T^5$. Let us again compare with the Hubble parameter

$$\frac{G^2 T^5}{H} \approx \frac{G^2 T^5}{\sqrt{G} T^2} = 1 \Rightarrow T^3 = \frac{1}{\sqrt{G}} = m_P^3. \quad (4.2.15)$$

Hence gravitons decouple already at the Planck energy at a time

$$t \approx \frac{1}{\sqrt{G}} \frac{1}{T^2} = \frac{1}{\sqrt{G}} \frac{1}{m_P^2} = \sqrt{G} = t_P, \quad (4.2.16)$$

which corresponds to the Planck time. Let us assume that even at such an early time we only have the particles of the standard model. This is probably unrealistic, and certainly wrong if the ideas about grand unified theories are correct. At the Planck time the entropy is given by

$$S = \frac{2\pi^2}{45} (g_G + g_*) (TR)^3 = \frac{2\pi^2}{45} \left(2 + \frac{427}{4}\right) (TR)^3. \quad (4.2.17)$$

If the graviton temperature today is T_G we have

$$T_G R' = TR. \quad (4.2.18)$$

Today the entropy is in gravitons, photons and neutrinos, such that

$$\begin{aligned} S &= \frac{2\pi^2}{45} (g_G (T_G R')^3 + g_\gamma (T_\gamma R')^3 + \frac{7}{8} (g_{\nu_e} + g_{\nu_\mu} + g_{\nu_\tau}) (T_\nu R')^3) \Rightarrow \\ &2(T_G R')^3 + 2(T_\gamma R')^3 + \frac{7}{8} 6(T_\gamma R')^3 \frac{4}{11} = \left(2 + \frac{427}{4}\right) (T_G R')^3 \Rightarrow \\ &\left(\frac{T_\gamma}{T_G}\right)^3 = \frac{427/4}{2 + 21/11} = \frac{4697}{172} \Rightarrow T_G = \left(\frac{172}{4697}\right)^{1/3} T_\gamma \Rightarrow T_G = 0.9\text{K}. \end{aligned} \quad (4.2.19)$$

Since at the Planck time there were probably more particles than the ones of the standard model, one should consider the above temperature as an upper limit. Measuring the temperature of the cosmic graviton background radiation would be very interesting, because it contains information about the number of relativistic degrees of freedom at the Planck time, and could hence be used to test ideas about grand unified theories. Unfortunately, it is totally unrealistic to hope for such an experimental result.

4.3 The End of Radiation Dominance

At about 1 min after the Big Bang, at temperatures in the MeV range, electrons and positrons have annihilated each other. After that only a few electrons

remained, which together with a few protons and neutrons form all the matter today. There were so few free charges that photons ultimately decoupled. At about 100000 years after the Big Bang, at temperatures in the eV range, the remaining electrons and protons (as well as other light atomic nuclei) combined into neutral atoms (mostly hydrogen). At this moment photon decoupling was completed. First, we will neglect the more complicated atomic nuclei, and assume that there are only electrons e , protons p and hydrogen atoms H . Due to charge neutrality we have

$$n_e = n_p, \quad (4.3.1)$$

and baryon number conservation implies

$$n_B = n_p + n_H. \quad (4.3.2)$$

In the temperature range we are interested in, electrons, protons and hydrogen atoms are non-relativistic. Assuming thermodynamical equilibrium we thus have

$$n_i = g_i \left(\frac{m_i T}{2\pi} \right)^{3/2} \exp(-\beta(m_i - \mu_i)), \quad i = e, p, H. \quad (4.3.3)$$

In chemical equilibrium the process $p + e \rightarrow H + \gamma$ implies

$$\mu_p + \mu_e = \mu_H. \quad (4.3.4)$$

Hence, we can write

$$\begin{aligned} n_H &= g_H \left(\frac{m_H T}{2\pi} \right)^{3/2} \exp(-\beta(m_H - \mu_p - \mu_e)) \\ &= \frac{g_H}{g_p g_e} n_p n_e \left(\frac{m_H}{m_p} \right)^{3/2} \left(\frac{m_e T}{2\pi} \right)^{-3/2} \exp(-\beta(m_H - m_p - m_e)) \\ &= n_p n_e \left(\frac{m_e T}{2\pi} \right)^{-3/2} \exp(\beta E_B). \end{aligned} \quad (4.3.5)$$

Here we have put $m_H = m_p$ and we have used $g_p = g_e = 2$ and $g_H = 4$. The hydrogen binding energy is

$$E_B = m_p + m_e - m_H = 13.6\text{eV}. \quad (4.3.6)$$

Let us consider the degree of ionization

$$\begin{aligned} X &= \frac{n_p}{n_B} = \frac{n_p}{n_p + n_H} \Rightarrow \\ \frac{1-X}{X^2} &= \frac{n_H(n_p + n_H)}{n_p^2} = n_B \left(\frac{m_e T}{2\pi} \right)^{-3/2} \exp(\beta E_B) \\ &= \frac{n_B}{n_\gamma} \frac{2}{\pi^2} \zeta(3) T^3 \left(\frac{m_e T}{2\pi} \right)^{-3/2} \exp(\beta E_B) \\ &= \frac{n_B}{n_\gamma} \frac{4\sqrt{2}}{\sqrt{\pi}} \zeta(3) \left(\frac{T}{m_e} \right)^{3/2} \exp(\beta E_B). \end{aligned} \quad (4.3.7)$$

This is the so-called Saha equation.

Let us assume that $\Omega_{baryons} = 0.1$, i.e. 10 percent of the critical energy density of the Universe is provided by baryonic matter. Then the Saha equation predicts that 90 percent of all electrons and protons will have formed hydrogen atoms at a temperature $T_R = 0.31$ eV. Here “R” stands for recombination, although electrons and protons had never been combined before. The above temperature corresponds to times of about 100000 years after the Bang. At that time the cosmic background radiation decouples from the matter, and the remaining ionization degree is frozen.

Finally, let us ask the question, when the era of radiation dominance ended, and matter dominated the Universe. Today the energy density of matter is

$$\rho_M = 1.88 \times 10^{-29} \Omega_0 \text{g cm}^{-3}. \quad (4.3.8)$$

Here Ω_0 is the matter contribution to the critical density. Photons and neutrinos carry the energy density

$$\rho_R = 8.09 \times 10^{-34} \text{g cm}^{-3}, \quad (4.3.9)$$

so obviously today the Universe is matter dominated. We know that $\rho_M \propto R^{-3}$ and $\rho_R \propto R^{-4}$, such that

$$\frac{\rho_R}{\rho_M} \propto R^{-1} \propto T, \quad (4.3.10)$$

where T is the temperature of radiation. What was that temperature when matter and radiation had the same energy density?

$$\frac{T}{2.7\text{K}} = \frac{1}{\rho_R/\rho_M} \Rightarrow T = 5.5\Omega_0\text{eV}. \quad (4.3.11)$$

The corresponding time is $t = 1.4 \times 10^3 \Omega^{-2}$ years.

Chapter 5

Nucleosynthesis

Today atomic nuclei are made in nuclear fusion inside stars. Via supernova explosions, heavy nuclei are distributed throughout the Universe, and that is also how they got to the earth. However, not all atomic nuclei — in particular not the light ones — have been made in that way. Before stars existed — namely in the early Universe — light atomic nuclei have been synthesized during the first few minutes after the Big Bang. Indeed nucleosynthesis provides a testing ground for our theories of the early Universe, because we can compare predicted abundances with observed ones. Both theoretical and observational quantities have uncertainties. How well do we know the nuclear reaction rates that govern primordial nucleosynthesis? How well can we determine the amount of a certain element (averaged over the whole Universe)? Are the abundances that we observe today the same that were produced in the Big Bang? Despite these uncertainties, the theory of nucleosynthesis provides an upper limit on the baryon density

$$\Omega_{baryons} \leq 0.18. \quad (5.0.1)$$

Hence, if we like to believe in inflation (and thus in $\Omega = 1$), we must assume that most of the energy in the Universe has not yet been identified (non-baryonic dark matter). Let us try to understand the arguments, that lead to the above limit.

5.1 The Neutron-Proton Ratio

Atomic nuclei consist of protons and neutrons. For the relative abundances of various nuclei the ratio of neutrons to protons is of central importance. The

binding energy of nuclei is in the MeV range, and hence nucleosynthesis takes place at temperatures in that range. Let us first consider very early times $t = 10^{-2}$ sec, corresponding to temperatures $T = 10$ MeV. At that time neutrinos were not yet decoupled, and protons and neutrons were in thermal equilibrium due to the weak interactions

$$p + e^- \rightarrow n + \nu_e, \quad n + e^+ \rightarrow p + \bar{\nu}_e. \quad (5.1.1)$$

Of course, at these times also some nuclei may have been present, and we want to estimate how many. Let us consider a nucleus of mass number A and mass m_A with Z protons and $A - Z$ neutrons. The corresponding density is then given by

$$n_A = g_A \left(\frac{m_A T}{2\pi} \right)^{3/2} \exp(-\beta(m_A - \mu_A)). \quad (5.1.2)$$

In “chemical” equilibrium we have

$$\mu_A = Z\mu_p + (A - Z)\mu_n. \quad (5.1.3)$$

Hence, we can write

$$\begin{aligned} n_A &= \frac{g_A}{g_p^Z g_n^{A-Z}} n_p^Z n_n^{A-Z} \left(\frac{m_A T}{2\pi} \right)^{3/2} \left(\frac{m_p T}{2\pi} \right)^{-3Z/2} \left(\frac{m_n T}{2\pi} \right)^{-3(A-Z)/2} \\ &\times \exp(-\beta(m_A - Zm_p - (A - Z)m_n)) \\ &= \frac{g_A}{2^A} n_p^Z n_n^{A-Z} A^{3/2} \left(\frac{m_p T}{2\pi} \right)^{-3(A-1)/2} \exp(\beta E_B). \end{aligned} \quad (5.1.4)$$

We have introduced the binding energy

$$E_B = Zm_p + (A - Z)m_n - m_A. \quad (5.1.5)$$

Let us now compute the fraction of mass that is in nuclei of type A (normalized to the total baryon number)

$$\begin{aligned} X_A &= \frac{n_A A}{n_B} \\ &= \frac{g_A}{2^A} X_p^Z X_n^{A-Z} n_B^{A-1} A^{3/2} \left(\frac{m_p T}{2\pi} \right)^{-3(A-1)/2} \exp(\beta E_B) \\ &= \frac{g_A}{2^A} \left(\zeta(3) \frac{2}{\pi^2} \right)^{A-1} A^{5/2} (2\pi)^{3(A-1)/2} \left(\frac{T}{m_p} \right)^{3(A-1)/2} \left(\frac{n_B}{n_\gamma} \right)^{A-1} X_p^Z X_n^{A-Z} \\ &\times \exp(\beta E_B). \end{aligned} \quad (5.1.6)$$

At this point we have reduced everything to the density of protons and neutrons. The neutron-proton ratio is given by

$$\frac{n_n}{n_p} = \exp(-\beta(m_n - m_p - \mu_n + \mu_p)) = \exp(-\beta(m_n - m_p - \mu_\nu + \mu_e)). \quad (5.1.7)$$

The neutron-proton mass difference is $m_n - m_p = 1.29$ MeV, while the chemical potentials are negligible compared to that. At a temperature of $T = 10$ MeV one then obtains

$$\begin{aligned} X_p \approx X_n \approx \frac{1}{2}, \quad X_2 = 6 \times 10^{-12}, \quad X_3 = 2 \times 10^{-23}, \quad X_4 = 2 \times 10^{-34}, \\ X_{12} = 2 \times 10^{-126}, \end{aligned} \quad (5.1.8)$$

i.e. the abundances of light nuclei are totally negligible.

Let us go one step further to temperatures at which the weak interactions go out of equilibrium, and the neutrinos decouple. As we have seen before, this happens around $T = 1$ MeV about 1 sec after the Bang. Then

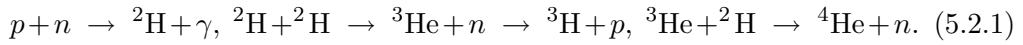
$$\begin{aligned} \frac{n_n}{n_p} &= \exp(-\beta(m_n - m_p)) \approx \frac{1}{6} \Rightarrow \\ X_p &\approx \frac{6}{7}, \quad X_n \approx \frac{1}{7}, \quad X_2 \approx 10^{-12}, \quad X_3 \approx 10^{-23}, \quad X_4 \approx 10^{-28}, \\ X_{12} &\approx 10^{-108}. \end{aligned} \quad (5.1.9)$$

From that time on, neutrons decay with a half-life of $\tau_n = 10.6(2)$ min. The remaining neutrons are finally turned into ${}^4\text{He}$, i.e. after nucleosynthesis

$$X_4 \approx \frac{4n_4}{n_B} = \frac{4n_n/2}{n_p + n_n} = \frac{2n_n/n_p}{1 + n_n/n_p} = \frac{2/6}{1 + 1/6} = \frac{2}{7} = 0.29. \quad (5.1.10)$$

5.2 Computed and Observed Abundances

About 1 min after the Big Bang nucleosynthesis begins, and it lasts for a few minutes. The whole thing happens at temperatures between 0.3 and 0.1 MeV. Here we cannot use equilibrium thermodynamics, since the rate of nuclear reactions is not strong enough to maintain an equilibrium. The most important reactions are



Direct reactions like $2p + 2n \rightarrow {}^4\text{He}$, on the other hand, are suppressed. Heavier nuclei than ${}^4\text{He}$ are difficult to make, because intermediate nuclei with $A = 5, 8$ are unstable. Also the Coulomb barrier for nuclear fusion is hard to overcome at low temperatures. At the end the mass ratios are computed by numerically integrating rate equations of the type

$$\frac{dX_A}{dt} = - \sum_i \lambda(A \rightarrow i) X_A + \sum_i \lambda(i \rightarrow A) X_i. \quad (5.2.2)$$

The result of these calculations depends on the ratio of baryon to photon density n_B/n_γ . The largest fraction of neutrons ends in ${}^4\text{He}$. The remaining protons later combine with electrons and become hydrogen. The fact that there are many more protons than neutrons is due to the intense cosmic background radiation. If it had not existed, many neutrons could have escaped their decay by going into deuterium, which now got photo-dissociated by the background radiation.

The ${}^4\text{He}$ abundance depends on the number of generations — to be precise on the number of light neutrino species. When there are many neutrino species, the Universe expands faster, and neutrinos decouple at a higher temperature. Therefore there are more neutrons and finally more ${}^4\text{He}$. The observed ${}^4\text{He}$ abundance leads to an upper bound of four generations. From LEP experiments of the Z_0 -width we know that there are only three generations (with light neutrinos). Still, it is interesting that cosmology also puts a bound on that quantity.

The ${}^2\text{H}$ abundance is very sensitive to the baryon density. At high density there are more reactions and hence more of the strongly bound ${}^4\text{He}$. At lower density more ${}^2\text{H}$ survives. The observed ${}^2\text{H}$ abundance puts an upper limit on the baryon density

$$\Omega_{\text{baryons}} \leq 0.18. \quad (5.2.3)$$

The measurements of abundances is a quite delicate matter. We are interested in the abundances immediately after the Big Bang, but we are observing today. Of course, today nucleosynthesis takes place in stars, and one must try to correct for that. At the end one measures the intensity of spectral lines in interstellar gas and derives the density of the various elements from that. One finds

$$\begin{aligned} X_4 &= 0.236(5), \quad X_2 = 2 \times 10^{-5} - 2 \times 10^{-4}, \quad X_3 = 2 \times 10^{-5} - 3 \times 10^{-4}, \\ X_7 &= 10^{-10} - 8 \times 10^{-10}. \end{aligned} \quad (5.2.4)$$

The fact that these abundances agree well with observation is a great success of the standard Big Bang theory. The physics after about 1 sec after the Bang seems to be well understood. We have seen how cosmological and particle physics questions are interrelated, for example, we have seen how cosmology can put a limit on the number of generations. Having discussed the physics in the MeV temperature range, we now want to go back to the even hotter and hence even earlier Universe. For that purpose we need the standard model of particle physics.

Chapter 6

The Standard Model of Particle Physics

The standard model of particle physics summarizes all we know about the fundamental forces of electromagnetism, as well as the weak and strong interactions. The standard model has been tested up to energies in the several hundred GeV range. When we want to talk about very early times ($t = 10^{-12}$ sec) in the early Universe, we must understand the physics of the standard model. Later we will speculate about even earlier times ($t = 10^{-34}$ sec), and we will do that based on grand unified theories, whose structure is very similar to that of the standard model. The standard model of particle physics is a relativistic quantum field theory, that combines the principles of quantum mechanics and special relativity. Like quantum electrodynamics (QED) the standard model is a gauge theory, however, with the non-Abelian gauge group $SU(3)_c \otimes SU(2)_L \otimes U(1)$. The gauge bosons are photons, W- and Z-bosons, as well as gluons. Gauge theories can exist in various phases: in the Coulomb phase with massless gauge bosons (like in QED), in the Higgs-phase with spontaneously broken gauge symmetry and with massive gauge bosons (as e.g. the W- and Z-bosons), and in the confinement phase, in which the gauge bosons do not appear in the spectrum (like the gluons in QCD). All these phases are realized in the standard model.

In particle physics symmetries play a central role. One distinguishes global and local symmetries. Global symmetries are usually only approximate. Exact symmetries, on the other hand, are locally realized, and require the existence of a gauge field. Our world is not as symmetric as the theories we use to describe it. This is because many symmetries are broken. The simplest form of symmetry

breaking is explicit breaking, which is due to non-invariant symmetry breaking terms in the classical Lagrangian of the theory. On the other hand, the quantization of the theory may also lead to explicit symmetry breaking, even if the classical Lagrangian is invariant. In that case one has an anomaly, that is due to an explicit symmetry breaking in the measure of the Feynman path integral. Only global symmetries are allowed to be explicitly broken (either in the Lagrangian or via an anomaly). Theories with explicitly broken gauge symmetries, on the other hand, are inconsistent (perturbatively non-renormalizable). For example, in the standard model all gauge anomalies are canceled, due to the properly arranged fermion content of each generation. For example, if the quarks would not have three colors, the Standard model would have a gauge anomaly and would thus be an inconsistent quantum field theory. A more interesting form of symmetry breaking is spontaneous symmetry breaking, which is a dynamical effect. When a global continuous symmetry breaks spontaneously, massless Goldstone bosons appear in the spectrum. If there is, in addition, a (weak) explicit symmetry breaking, the Goldstone bosons pick up a (small) mass. When a gauge symmetry is spontaneously broken, one has the so-called Higgs mechanism, which gives mass to the gauge bosons. This gives rise to an additional helicity state. This state has the quantum numbers of a Goldstone bosons, if the symmetry were global. One says that the gauge bosons eat the Goldstone bosons, and thus become massive.

6.1 Scalar Electrodynamics

The proto-type of a gauge theory is quantum electrodynamics, the theory of the electromagnetic interaction between charged electrons and positrons via photon exchange. We want to make our life easy, and consider charged particles without spin — so-called scalars. For example, we can think of the Cooper pairs in a superconductor. A charged scalar particle is described by a complex field $\Phi(x) \in \mathbb{C}$. One needs two real degrees of freedom, in order to describe particle and anti-particle. The space-time point x is in a flat Minkowski space, because we don't know how to include gravity in a quantum field theory. As we will discuss later, a quantum field theory is defined by a Feynman path integral over all field configurations

$$Z = \int \mathcal{D}\Phi \exp(-\frac{i}{\hbar} S[\Phi]). \quad (6.1.1)$$

Here $S[\Phi]$ is the classical action of the field Φ . Let us go back to classical mechanics of point particles for a moment. The solutions of the theory are classical

paths $x(t)$, and the action is a functional of a path

$$S[x] = \int dt L(x, \partial_t x), \quad (6.1.2)$$

where the Lagrange function is given by

$$L(x, \partial_t x) = \frac{m}{2} \partial_t x \partial_t x - V(x). \quad (6.1.3)$$

The classical equation of motion follows from the variational principle of the action

$$\partial_t \frac{\delta L}{\delta \partial_t x} - \frac{\delta L}{\delta x} = m \partial_t^2 x + \nabla V(x) = 0. \quad (6.1.4)$$

This is just Newton's equation of motion. There is a complete analogy with field theory. The solutions of classical field theory are field configurations $\Phi(x) = \Phi_1(x) + i\Phi_2(x)$, and the action is a functional of the field configuration

$$S[\Phi] = \int d^4x \mathcal{L}(\Phi, \partial_\mu \Phi). \quad (6.1.5)$$

Here we are dealing with a Lagrange density (or Lagrangian)

$$\mathcal{L}(\Phi, \partial_\mu \Phi) = \partial^\mu \Phi^* \partial_\mu \Phi - V(\Phi) = \partial^\mu \Phi_1 \partial_\mu \Phi_1 + \partial^\mu \Phi_2 \partial_\mu \Phi_2 - V(\Phi_1, \Phi_2). \quad (6.1.6)$$

The simplest form of a potential is the harmonic oscillator

$$V(\Phi) = m^2 \Phi^* \Phi = m^2 |\Phi|^2 = m^2 (\Phi_1^2 + \Phi_2^2). \quad (6.1.7)$$

The classical field equations then take the form

$$\partial_\mu \frac{\delta \mathcal{L}}{\delta \partial_\mu \Phi_i} - \frac{\delta \mathcal{L}}{\delta \Phi_i} = \partial_\mu \partial^\mu \Phi_i + m^2 \Phi_i = 0. \quad (6.1.8)$$

This is the 2-component Klein-Gordon equation for a free scalar field. The Lagrange density has a symmetry. It is invariant under global $U(1)$ transformations

$$\Phi(x)' = \exp(ie\varphi)\Phi(x) \Rightarrow \Phi^*(x)' = \exp(-ie\varphi)\Phi^*(x), \quad (6.1.9)$$

because then

$$\partial_\mu \Phi(x)' = \exp(ie\varphi) \partial_\mu \Phi(x), \quad \partial_\mu \Phi^*(x)' = \exp(-ie\varphi) \partial_\mu \Phi^*(x). \quad (6.1.10)$$

This invariance is related to the conservation of the charge e . We now also want to allow interactions by generalizing the potential $V(\Phi)$. Using the $U(1)$ invariance as a guiding principle, we come to a $|\Phi|^4$ theory

$$V(\Phi) = m^2 |\Phi|^2 + \lambda |\Phi|^4. \quad (6.1.11)$$

The coupling constant λ must be positive, in order for the potential to be bounded from below. One can, however, choose $m^2 < 0$. Hence, we distinguish two cases. For $m^2 > 0$ the potential has a single minimum at $\Phi = 0$. The classical solution of lowest energy (the classical vacuum) is simply the constant field $\Phi(x) = 0$. When $m^2 < 0$, this trivial vacuum configuration is unstable, because it corresponds to a maximum of the potential. The condition for a minimum now reads

$$\frac{\partial V}{\partial \Phi_i} = 2m^2 \Phi_i + 4\lambda |\Phi|^2 \Phi_i = 0 \Rightarrow |\Phi|^2 = -\frac{m^2}{2\lambda}. \quad (6.1.12)$$

The true vacuum is no longer unique. Instead, there is a whole class of degenerate vacua

$$\Phi(x) = v \exp(i\chi), \quad v = \sqrt{-\frac{m^2}{4\lambda}}, \quad (6.1.13)$$

parameterized by an angle $\chi \in [0, 2\pi[$. The quantity v is the vacuum expectation value of the field Φ . Let us choose the vacuum state with $\chi = 0$ and let us expand around the corresponding minimum

$$\begin{aligned} \Phi &= v + \sigma + i\pi \Rightarrow \Phi^* = v + \sigma - i\pi, \\ \partial_\mu \Phi &= \partial_\mu \sigma + i\partial_\mu \pi, \quad \partial^\mu \Phi^* = \partial^\mu \sigma - i\partial^\mu \pi, \\ |\Phi|^2 &= (v + \sigma)^2 + \pi^2 \Rightarrow \\ \mathcal{L}(\sigma, \partial_\mu \sigma, \pi, \partial_\mu \pi) &= \partial^\mu \sigma \partial_\mu \sigma + \partial^\mu \pi \partial_\mu \pi - m^2(v + \sigma)^2 - m^2 \pi^2 \\ &\quad - \lambda((v + \sigma)^2 + \pi^2)^2 \\ &\approx \partial^\mu \sigma \partial_\mu \sigma + \partial^\mu \pi \partial_\mu \pi - m^2 v^2 - 2m^2 v \sigma - m^2 \sigma^2 - m^2 \pi^2 \\ &\quad - \lambda(v^4 + 4v^3 \sigma + 6v^2 \sigma^2 + 2v^2 \pi^2) \\ &= \partial^\mu \sigma \partial_\mu \sigma - (m^2 + 6\lambda v^2) \sigma^2 + \partial^\mu \pi \partial_\mu \pi. \end{aligned} \quad (6.1.14)$$

One finds a σ -particle of mass

$$m_\sigma^2 = m^2 + 6\lambda v^2 = m^2 - 6\lambda \frac{m^2}{2\lambda} = -2m^2 > 0, \quad (6.1.15)$$

as well as a massless π -particle $m_\pi = 0$. This massless particle is a so-called Goldstone boson. Its presence is characteristic for the spontaneous breaking of the global $U(1)$ symmetry.

Now we want to promote the global $U(1)$ symmetry to a local one. We demand $U(1)$ gauge invariance

$$\Phi(x)' = \exp(ie\varphi(x))\Phi(x), \quad (6.1.16)$$

where $\varphi(x)$ now is a space-time dependent transformation parameter. The potential is gauge invariant ($V(\Phi') = V(\Phi)$) because

$$|\Phi(x)'|^2 = \Phi^*(x)' \Phi(x)' = \Phi^*(x) \exp(-ie\varphi(x)) \exp(ie\varphi(x)) \Phi(x) = |\Phi(x)|^2. \quad (6.1.17)$$

The kinetic term, on the other hand, is not invariant, because

$$\partial_\mu \Phi(x)' = \exp(ie\varphi(x))(\partial_\mu \Phi(x) + ie\partial_\mu \varphi(x)\Phi(x)). \quad (6.1.18)$$

We introduce a gauge field $A_\mu(x)$, that transforms such that the $\partial_\mu \varphi$ term is eliminated, i.e.

$$A_\mu(x)' = A_\mu(x) + \partial_\mu \varphi(x). \quad (6.1.19)$$

Then

$$(\partial_\mu - ieA_\mu(x)')\Phi(x)' = \exp(ie\varphi(x))(\partial_\mu - ieA_\mu(x))\Phi(x) \quad (6.1.20)$$

is gauge covariant. The above quantity is the covariant derivative

$$D_\mu \Phi(x) = (\partial_\mu - ieA_\mu(x))\Phi(x). \quad (6.1.21)$$

The Lagrange function can now be written in a gauge invariant way

$$\mathcal{L}(\Phi, \partial_\mu \Phi, A_\mu) = D^\mu \Phi^* D_\mu \Phi - V(\Phi). \quad (6.1.22)$$

Up to now, the gauge field A_μ appeared only as an external field. We have not yet written a kinetic term for it. From classical electrodynamics we know how to write such a term. We construct the field strength tensor

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu, \quad (6.1.23)$$

which is gauge invariant

$$F'_{\mu\nu} = \partial_\mu A'_\nu - \partial_\nu A'_\mu = \partial_\mu A_\nu + \partial_\mu \partial_\nu \varphi - \partial_\nu A_\mu - \partial_\nu \partial_\mu \varphi = F_{\mu\nu}. \quad (6.1.24)$$

The Lagrange density of the free electromagnetic field is

$$\mathcal{L}(A_\mu) = -\frac{1}{4}F^{\mu\nu}F_{\mu\nu}, \quad (6.1.25)$$

which leads to Maxwell's equations

$$\partial^\mu F_{\mu\nu} = 0. \quad (6.1.26)$$

The total Lagrange function of scalar QED takes the form

$$\mathcal{L}(\Phi, \partial_\mu \Phi, A_\mu) = -\frac{1}{4}F^{\mu\nu}F_{\mu\nu} + D^\mu \Phi^* D_\mu \Phi - V(\Phi). \quad (6.1.27)$$

A mass term $m_\gamma^2 A^\mu A_\mu$ is not allowed, because it would violate gauge invariance.

Like for the global $U(1)$ symmetry, we distinguish two cases. For $m^2 > 0$ the symmetry is unbroken, and we have a Coulomb phase with scalar particles of

charge e and massless photons. The broken phase at $m^2 < 0$ is more interesting. Again, there are degenerate vacua, but they are now related by gauge transformations. Hence, it is useful to fix the gauge. We choose a physical unitary gauge

$$\text{Im}\Phi(x) = \Phi_2(x) = 0, \text{Re}\Phi(x) = \Phi_1(x) \geq 0. \quad (6.1.28)$$

Let us again consider fluctuation around the vacuum value v

$$\Phi = v + \sigma. \quad (6.1.29)$$

Now there is no π excitation, because in the unitary gauge the field Φ is real and positive. We find

$$\begin{aligned} V(\Phi) &= m^2(v + \sigma)^2 + \lambda(v + \sigma)^4 \\ &= m^2v^2 + 2m^2v\sigma + m^2\sigma^2 + \lambda(v^4 + 4v^3\sigma + 6v^2\sigma^2) + \dots \\ &= (m^2 + 6\lambda v^2)\sigma^2 + \dots = -2m^2\sigma^2 + \dots, \end{aligned} \quad (6.1.30)$$

i.e., again there is a massive σ particle, but no longer a massless π . What happened to this degree of freedom? Let us consider the gauge field

$$\begin{aligned} D^\mu \Phi^* D_\mu \Phi &= (\partial^\mu + ieA^\mu)(v + \sigma)(\partial_\mu - ieA_\mu)(v + \sigma) \\ &= (\partial^\mu \sigma + ieA^\mu v + ieA^\mu \sigma)(\partial_\mu \sigma - ieA_\mu v - ieA_\mu \sigma) \\ &= \partial^\mu \sigma \partial_\mu \sigma + e^2 v^2 A^\mu A_\mu + \dots \end{aligned} \quad (6.1.31)$$

We obtain a massive photon

$$m_\gamma^2 = ev. \quad (6.1.32)$$

This mechanism of mass generation is called the Higgs mechanism. It is based on the spontaneous breakdown of the gauge symmetry. A phase, in which the gauge symmetry is spontaneously broken, and in which the gauge bosons hence are massive, is called a Higgs phase. In our Universe, the $U(1)$ gauge symmetry of electrodynamics is unbroken, the photon is massless, and we are in a Coulomb phase. Still, intelligent human beings can create the coolest spot of the Universe here on earth and generate superconductivity in condensed matter. Then the $U(1)$ gauge symmetry indeed gets spontaneously broken, and the photon becomes massive. The mass can be measured, because it is related to the penetration depth of magnetic fields in the superconductor.

6.2 The Electroweak Interaction

The electroweak interaction is described by an $SU(2) \otimes U(1)$ gauge theory, whose symmetry gets spontaneously broken to the $U(1)$ of electromagnetism. Again, a

scalar Higgs field Φ plays a central role in the dynamics. However, the field Φ now has two complex components (complex doublet)

$$\Phi(x) = \begin{pmatrix} \Phi_+(x) \\ \Phi_0(x) \end{pmatrix}, \quad \Phi_+(x), \Phi_0(x) \in \mathbb{C}. \quad (6.2.1)$$

The indexes $+$ and 0 will turn out to denote electric charges. Again, let us first discuss a model with global symmetry

$$\mathcal{L}(\Phi, \partial_\mu \Phi) = \partial^\mu \Phi^\dagger \partial_\mu \Phi - V(\Phi), \quad (6.2.2)$$

where

$$V(\Phi) = m^2 |\Phi|^2 + \lambda |\Phi|^4. \quad (6.2.3)$$

Of course, now

$$|\Phi|^2 = \Phi^\dagger \Phi = \Phi_+^* \Phi_+ + \Phi_0^* \Phi_0 = \Phi_{+1}^* \Phi_{+1} + \Phi_{+2}^* \Phi_{+2} + \Phi_{01}^* \Phi_{01} + \Phi_{02}^* \Phi_{02}. \quad (6.2.4)$$

The Lagrange density is invariant under $SU(2)$ transformations.

$$\Phi(x)' = g\Phi(x), \quad g \in SU(2). \quad (6.2.5)$$

$SU(2)$ is the group of unitary 2×2 matrices with determinant 1

$$g^\dagger g = gg^\dagger = 1, \quad g^\dagger = g^{T*}, \quad \det g = 1. \quad (6.2.6)$$

A general $SU(2)$ matrix can be written as

$$\begin{aligned} g &= \begin{pmatrix} g_1 & -g_2^* \\ g_2 & g_1^* \end{pmatrix} \Rightarrow g^\dagger = \begin{pmatrix} g_1^* & g_2^* \\ -g_2 & g_1 \end{pmatrix} \Rightarrow \\ gg^\dagger &= \begin{pmatrix} |g_1|^2 + |g_2|^2 & 0 \\ 0 & |g_1|^2 + |g_2|^2 \end{pmatrix} = 1 \Rightarrow \\ |g_1|^2 + |g_2|^2 &= 1, \quad \det g = |g_1|^2 + |g_2|^2 = 1. \end{aligned} \quad (6.2.7)$$

The space of $SU(2)$ matrices is isomorphic to the 3-dimensional sphere S^3 . The global $SU(2)$ invariance of the above Lagrange density follows from

$$\begin{aligned} |\Phi'|^2 &= \Phi^\dagger(x)' \Phi(x)' = (g\Phi(x))^\dagger g\Phi(x) = \Phi^\dagger(x) g^\dagger g \Phi(x) = |\Phi|^2, \\ \partial^\mu \Phi^\dagger(x)' \partial_\mu \Phi(x)' &= \partial^\mu \Phi^\dagger(x) g^\dagger g \partial_\mu \Phi(x) = \partial^\mu \Phi^\dagger(x) \partial_\mu \Phi(x). \end{aligned} \quad (6.2.8)$$

In fact, the Lagrange density is invariant even under $O(4) = SU(2) \otimes SU(2)$ transformations. However, only one $SU(2)$ symmetry is gauged in the standard model.

Again, we must distinguish between the symmetric phase at $m^2 > 0$ and the broken phase at $m^2 < 0$. For $m^2 > 0$ there is a unique vacuum state

$$\Phi(x) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \quad (6.2.9)$$

For $m^2 < 0$ the vacuum is degenerate and we must make our choice

$$\Phi(x) = \begin{pmatrix} 0 \\ v \end{pmatrix}, \quad v \in \mathbb{R}_{>0}. \quad (6.2.10)$$

Again, we expand around the vacuum state

$$\Phi(x) = \begin{pmatrix} \pi_1 + i\pi_2 \\ v + \sigma + i\pi_3 \end{pmatrix}, \quad (6.2.11)$$

and we obtain

$$\begin{aligned} \partial^\mu \Phi^\dagger \partial_\mu \Phi &= \partial^\mu \sigma \partial_\mu \sigma + \partial^\mu \pi_1 \partial_\mu \pi_1 + \partial^\mu \pi_2 \partial_\mu \pi_2 + \partial^\mu \pi_3 \partial_\mu \pi_3, \\ V(\Phi) &= m^2((v + \sigma)^2 + \pi_1^2 + \pi_2^2 + \pi_3^2) + \lambda((v + \sigma)^2 + \pi_1^2 + \pi_2^2 + \pi_3^2)^2 \\ &= m^2(v^2 + 2v\sigma + \sigma^2 + \pi_1^2 + \pi_2^2 + \pi_3^2) \\ &\quad + \lambda(v^4 + 4v^3\sigma + 6v^2\sigma^2 + 2v^2(\pi_1^2 + \pi_2^2 + \pi_3^2)) + \dots \\ &= -2m^2\sigma^2 + \dots \end{aligned} \quad (6.2.12)$$

Again, we find a σ particle of mass

$$m_\sigma^2 = -2m^2, \quad (6.2.13)$$

and in this case three massless Goldstone bosons π_1, π_2, π_3 .

As before, we want to promote the global $SU(2)$ symmetry to a local one, i.e. we demand symmetry under

$$\Phi(x)' = g(x)\Phi(x). \quad (6.2.14)$$

The potential $V(\Phi)$ is trivially gauge invariant. The kinetic term, on the other hand, is not invariant, because

$$\partial_\mu \Phi(x)' = g(x)\partial_\mu \Phi(x) + \partial_\mu g(x)\Phi(x) = g(x)(\partial_\mu \Phi(x) + g^\dagger(x)\partial_\mu g(x)\Phi(x)). \quad (6.2.15)$$

Again, we want to compensate the additional term by a gauge field. We introduce

$$W_\mu(x)' = g(x)(W_\mu(x) + \partial_\mu)g^\dagger(x), \quad (6.2.16)$$

where $W_\mu(x)$ is an anti-Hermitean matrix, that can be written as

$$W_\mu(x) = igW_\mu^a(x)\sigma_a, \quad a = 1, 2, 3. \quad (6.2.17)$$

The covariant derivative takes the form

$$D_\mu\Phi(x) = (\partial_\mu + W_\mu(x))\Phi(x), \quad (6.2.18)$$

and one finds

$$\begin{aligned} D_\mu\Phi(x)' &= (\partial_\mu + W_\mu(x)')\Phi(x)' \\ &= g(x)[\partial_\mu\Phi(x) + g^\dagger(x)\partial_\mu g(x)\Phi(x) \\ &\quad + W_\mu(x)g^\dagger(x)g(x)\Phi(x) + \partial_\mu g^\dagger(x)g(x)\Phi(x)] \\ &= g(x)(\partial_\mu W_\mu(x))\Phi(x) = g(x)D_\mu\Phi(x). \end{aligned} \quad (6.2.19)$$

As before, we introduce the gauge invariant Lagrange density

$$\mathcal{L}(\Phi, \partial_\mu\Phi, W_\mu) = D^\mu\Phi^\dagger D_\mu\Phi - V(\Phi). \quad (6.2.20)$$

Here the gauge field is an external field. We still have to add the kinetic term. The field strength of a non-Abelian gauge field is given by

$$W_{\mu\nu} = \partial_\mu W_\nu - \partial_\nu W_\mu + [W_\mu, W_\nu], \quad (6.2.21)$$

and it transforms as

$$W'_{\mu\nu} = gW_{\mu\nu}g^\dagger. \quad (6.2.22)$$

In analogy with Abelian gauge theory we write

$$\mathcal{L}(W_\mu) = -\frac{1}{4}\text{Tr}W^{\mu\nu}W_{\mu\nu}, \quad (6.2.23)$$

which is gauge invariant. In contrast to Abelian gauge fields, non-Abelian gauge fields are themselves charged, and hence interact even without other charged matter fields present.

Thus far, we have limited ourselves to transformations with determinant 1. Now we want to discuss the $U(1)$ transformations related to the determinant. The scalar field then transforms as

$$\Phi(x)' = \exp(ig'\varphi(x))\Phi(x). \quad (6.2.24)$$

Here g' is a new coupling constant — the weak hypercharge. The corresponding $U(1)$ gauge field transforms as

$$B_\mu(x)' = B_\mu(x) + \partial_\mu\varphi(x). \quad (6.2.25)$$

The new field yields an additional term to the covariant derivative

$$\begin{aligned} D_\mu \Phi(x) &= (\partial_\mu + W_\mu(x) - ig'B_\mu(x))\Phi(x) \\ &= (\partial_\mu + igW_\mu^a(x)\sigma_a - ig'B_\mu(x)) \begin{pmatrix} \Phi_+(x) \\ \Phi_0(x) \end{pmatrix}. \end{aligned} \quad (6.2.26)$$

Once more, we introduce a gauge invariant field strength

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu, \quad (6.2.27)$$

and we obtain the total Lagrange density

$$\mathcal{L}(\Phi, W_\mu, B_\mu) = D^\mu \Phi^\dagger D_\mu \Phi - V(\Phi) - \frac{1}{4} \text{Tr} W^{\mu\nu} W_{\mu\nu} - \frac{1}{4} B^{\mu\nu} B_{\mu\nu}. \quad (6.2.28)$$

Let us consider the symmetry breaking case $m^2 < 0$, again in the unitary gauge

$$\Phi(x) = \begin{pmatrix} 0 \\ v \end{pmatrix}, \quad v \in \mathbb{R}_{>0}. \quad (6.2.29)$$

This vacuum state is still invariant under $U(1)$ gauge transformations of the type

$$\Phi(x)' = \begin{pmatrix} \exp(ie\varphi(x)) & 0 \\ 0 & 1 \end{pmatrix} \Phi(x), \quad (6.2.30)$$

which have a $U(1)_Y$ hypercharge part, but also a diagonal $SU(2)_L$ part. Hence, the vacuum state does not break $SU(2)_L \otimes U(1)_Y$ symmetry completely. Instead, there is a remaining $U(1)$ symmetry, that we will soon identify with the one of electromagnetism. Since that symmetry remains unbroken, despite the Higgs mechanism, there will be one massless gauge boson — the photon. All other gauge bosons eat a Goldstone bosons and become massive. We consider the fluctuations in unitary gauge

$$\Phi(x) = \begin{pmatrix} 0 \\ v + \sigma \end{pmatrix}. \quad (6.2.31)$$

Expanding in powers of the real field σ , we obtain

$$\begin{aligned} D^\mu \Phi^\dagger D_\mu \Phi &= |(\partial_\mu + igW_\mu^a \sigma_a - ig'B_\mu) \begin{pmatrix} 0 \\ v + \sigma \end{pmatrix}|^2 \\ &= \partial^\mu \sigma \partial_\mu \sigma + (v + \sigma)^2 (01)(gW^{\mu a} \sigma_a - g'B^\mu)(gW_\mu^a \sigma_a - g'B_\mu) \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= \partial^\mu \sigma \partial_\mu \sigma + (v + \sigma)^2 (g^2 W^{\mu 1} W_\mu^1 + g^2 W^{\mu 2} W_\mu^2 \\ &\quad + (gW^{\mu 3} + g'B^\mu)(gW_\mu^3 + g'B_\mu)). \end{aligned} \quad (6.2.32)$$

Also we have

$$V(\Phi) = m^2(v + \sigma)^2 + \lambda(v + \sigma)^4 = -2m^2\sigma^2 + \dots, \quad (6.2.33)$$

and hence there is a Higgs particle of mass

$$m_\sigma^2 = -2m^2. \quad (6.2.34)$$

Also, there are two W-bosons of mass

$$m_W = gv. \quad (6.2.35)$$

Furthermore, we introduce the linear combination

$$Z_\mu = \frac{gW_\mu^3 + g'B_\mu}{\sqrt{g^2 + g'^2}}, \quad (6.2.36)$$

which has the mass

$$m_Z = \sqrt{g^2 + g'^2}v. \quad (6.2.37)$$

The other linear independent combination

$$A_\mu = \frac{gB_\mu - g'W_\mu^3}{\sqrt{g^2 + g'^2}}, \quad (6.2.38)$$

remains massless and describes the photon. We introduce the Weinberg angle θ_W via

$$\begin{pmatrix} A_\mu \\ Z_\mu \end{pmatrix} = \begin{pmatrix} \cos \theta_W & -\sin \theta_W \\ \sin \theta_W & \cos \theta_W \end{pmatrix} \begin{pmatrix} B_\mu \\ W_\mu^3 \end{pmatrix}, \quad (6.2.39)$$

such that

$$\frac{g}{\sqrt{g^2 + g'^2}} = \cos \theta_W, \quad \frac{g'}{\sqrt{g^2 + g'^2}} = \sin \theta_W, \quad (6.2.40)$$

and hence

$$\frac{m_Z}{m_W} = \frac{g}{\sqrt{g^2 + g'^2}} = \cos \theta_W. \quad (6.2.41)$$

The experimental values for the masses are $m_W = 80$ GeV and $m_Z = 91$ GeV, such that $\sin^2 \theta_W = 0.229$. The coupling constant of the photon is the charge e .

On the other hand, the corresponding covariant derivative of the scalar field is

$$\begin{aligned}
D_\mu \Phi &= (\partial_\mu + igW_\mu^1 \sigma_1 + igW_\mu^2 \sigma_2 + igW_\mu^3 \sigma_3 - ig'B_\mu) \begin{pmatrix} \Phi_+ \\ \Phi_0 \end{pmatrix} \\
&= (\partial_\mu + igW_\mu^1 \sigma_1 + igW_\mu^2 \sigma_2 \\
&\quad + i \begin{pmatrix} gW_\mu^3 - g'B_\mu & 0 \\ 0 & -gW_\mu^3 - g'B_\mu \end{pmatrix}) \begin{pmatrix} \Phi_+ \\ \Phi_0 \end{pmatrix} \\
&= (\partial_\mu + igW_\mu^1 \sigma_1 + igW_\mu^2 \sigma_2 \\
&\quad + i \begin{pmatrix} \frac{(g^2 - g'^2)Z_\mu}{\sqrt{g^2 + g'^2}} - \frac{2gg'A_\mu}{\sqrt{g^2 + g'^2}} & 0 \\ 0 & \sqrt{g^2 + g'^2}Z_\mu \end{pmatrix}) \begin{pmatrix} \Phi_+ \\ \Phi_0 \end{pmatrix}. \quad (6.2.42)
\end{aligned}$$

We identify the electric charge as

$$e = \frac{2gg'}{\sqrt{g^2 + g'^2}}. \quad (6.2.43)$$

Now we see that indeed only Φ_+ couples to the electric field. It has charge e , while Φ_0 is neutral.

6.3 The Leptons

The leptons participate in the electroweak, but not in the strong interactions. They are fermions coming in three generations: the electron and its neutrino, the muon and its neutrino, and the τ and its neutrino. In the standard model the neutrinos are massless chiral fermions with only one helicity state. There are left-handed neutrinos and right-handed anti-neutrinos only. The left-handed neutrinos and the left-handed components of the charged fermions form $SU(2)_L$ doublets

$$\begin{pmatrix} \nu_{eL}(x) \\ e_L(x) \end{pmatrix}, \begin{pmatrix} \nu_{\mu L}(x) \\ \mu_L(x) \end{pmatrix}, \begin{pmatrix} \nu_{\tau L}(x) \\ \tau_L(x) \end{pmatrix}, \quad (6.3.1)$$

while the right-handed components of the charged fermions $e_R(x)$, $\mu_R(x)$ and $\tau_R(x)$ are $SU(2)$ singlets. The left-handed doublets carry hypercharge $-g'$ and the right-handed singlets carry $-2g'$. Considering the coupling to A_μ one then finds that the neutrinos are indeed neutral, while e , μ and τ carry the charge e . The Lagrange function of the leptons of the first generation takes the form

$$\begin{aligned}
\mathcal{L}(e_L, \nu_{eL}, e_R, W_\mu, B_\mu) &= (\bar{\nu}_{eL} \bar{e}_L) i\gamma^\mu (\partial_\mu + igW_\mu^a \sigma_a + ig'B_\mu) \begin{pmatrix} \nu_{eL} \\ e_L \end{pmatrix} \\
&\quad + \bar{e}_R i\gamma^\mu (\partial_\mu + 2ig'B_\mu) e_R. \quad (6.3.2)
\end{aligned}$$

So far we have not introduced any mass terms for the leptons. An electron mass term would have the form $m_e[\bar{e}_R e_L + \bar{e}_L e_R]$. Such a term is not gauge invariant, because the left- and right-handed components transform differently under $SU(2)_L$. Hence, explicit mass terms are forbidden in chiral gauge theories. The scalar field Φ gave mass to the gauge bosons via spontaneous symmetry breaking. Similarly, it can give mass to the fermions. Let us write down a Yukawa coupling

$$\mathcal{L}(e_L, \nu_{eL}, e_R, \Phi) = f_e[\bar{e}_R(\Phi_+^* \Phi_0^*) \begin{pmatrix} \nu_{eL} \\ e_L \end{pmatrix} + (\bar{\nu}_{eL} \bar{e}_L \begin{pmatrix} \Phi_+ \\ \Phi_0 \end{pmatrix} e_R)]. \quad (6.3.3)$$

Again, using the vacuum state from before, we obtain

$$\mathcal{L}(e_L, \nu_{eL}, e_R, \Phi) = f_e[\bar{e}_R(0v) \begin{pmatrix} \nu_{eL} \\ e_L \end{pmatrix} + (\bar{\nu}_{eL} \bar{e}_L \begin{pmatrix} 0 \\ v \end{pmatrix} e_R)] = f_e v[\bar{e}_R e_L + \bar{e}_L e_R]. \quad (6.3.4)$$

Hence the electrons obtain the mass

$$m_e = f_e v, \quad (6.3.5)$$

while the neutrinos indeed remain massless.

6.4 The Strong Interactions

QCD is a relativistic quantum field theory with a non-Abelian $SU(3)_C$ gauge symmetry, describing the interaction between quarks via gluon exchange. QCD is formulated in terms of quark and gluon fields. Still, the observed particles are not directly quanta of these fields — namely quarks and gluons — but hadrons (confined multi-quark and gluon states). This means that a perturbative treatment of QCD with free quarks and gluons as in and out states is inappropriate. Confinement is a complicated nonperturbative phenomenon that could not yet be derived from QCD first principles. A nonperturbative formulation of QCD is provided by lattice gauge theory. Its numerical treatment (Monte-Carlo simulation) has confirmed various nonperturbative phenomena — among them confinement. In recent years the results have become more quantitative, but still QCD is far from being well tested nonperturbatively.

Fortunately, there are aspects of the strong interactions that can be understood using QCD perturbation theory. This is due to the property of asymptotic freedom — the fact that quarks and gluons behave like free particles at high energies. In this way one can derive the parton model from QCD, identifying

the partons with quarks and gluons. Of course, quarks and gluons do interact as long as the energy is finite. We introduce the quarks as elementary fermions with flavor and color quantum numbers described by a Dirac spinor field $\Psi(x)$. The Dirac equation for a hypothetical free quark with flavor $f \in \{u, d, s, c, b, t, \}$ and color $c \in \{r, g, b\}$ is given by

$$(i\gamma^\mu \partial_\mu - m_f)\Psi_{fc}(x) = 0. \quad (6.4.1)$$

Here m_f is the flavor dependent (but color independent) quark mass. The Dirac equation follows from the Lagrange density

$$\mathcal{L}(\bar{\Psi}\Psi) = \sum_{f,c} \bar{\Psi}_{fc}(x)(i\gamma^\mu \partial_\mu - m_f)\Psi_{fc}(x) = \sum_c \bar{\Psi}_c(x)(i\gamma^\mu \partial_\mu - \mathcal{M})\Psi_c(x). \quad (6.4.2)$$

Here we have introduced the quark mass matrix

$$\mathcal{M} = \text{diag}(m_u, m_d, m_s, m_c, m_b, m_t). \quad (6.4.3)$$

Since quarks of different color have the same mass, the above Lagrange function has a large symmetry. We can rotate each flavor f by a matrix U^f in color space

$$\Psi'_{fc'}(x) = U^f_{c'c} \Psi_{fc}(x) \text{ or } \Psi'_f(x) = U^f \Psi_f(x). \quad (6.4.4)$$

One has

$$\bar{\Psi}_f(x) = \Psi_f^\dagger(x)\gamma^0, \quad (6.4.5)$$

and hence

$$\bar{\Psi}'_f(x) = \Psi'^{\dagger}_f(x)\gamma^0 = [U^f \Psi_f(x)]^\dagger \gamma^0 = \Psi_f^\dagger(x) U^{f\dagger} \gamma^0 = \bar{\Psi}_f(x) U^{f\dagger}. \quad (6.4.6)$$

The Lagrange function remains invariant if

$$U^{f\dagger} U^f = 1 \Rightarrow U^f \in U(3). \quad (6.4.7)$$

Each flavor can be rotated independently. Hence the symmetry of the free Lagrange function is

$$\otimes_f U(3)_f = U(3)_u \otimes U(3)_d \otimes U(3)_s \otimes U(3)_c \otimes U(3)_b \otimes U(3)_t \quad (6.4.8)$$

The weak interactions mix the flavors, such that they can no longer be rotated independently. In fact, they break the above symmetry explicitly down to

$$\otimes_f U(3)_f \rightarrow U(3) \otimes U(1)_{em}. \quad (6.4.9)$$

The remaining $U(3)$ symmetry rotates all flavors in the same way

$$\Psi'_f(x) = U \Psi_f(x). \quad (6.4.10)$$

$U(1)_{em}$ is the symmetry of electromagnetism, it rotates the quark spinors by a complex phase depending on their electric charge. The remaining $U(3)$ symmetry can be decomposed in a $U(1)_B$ and an $SU(3)_C$ symmetry

$$U(3) = U(1)_B \otimes SU(3)_C. \quad (6.4.11)$$

A $U(1)_B$ transformation multiplies all quark spinor by the same phase

$$\Psi'_{fc}(x) = \exp(i\varphi_B)\Psi_{fc}(x). \quad (6.4.12)$$

The invariance of the free Lagrange function leads to a conserved quantum number — the baryon number. When we add the weak interactions, the classical Lagrange density is still $U(1)_B$ invariant. However, baryon number is not conserved in the standard model, because it does not survive the quantization of the theory. In fact, there is an anomaly in the $U(1)_B$ symmetry, which corresponds to an explicit symmetry breaking. The breaking becomes appreciable only at very high energies in the TeV range, but would then lead to baryon number violating processes in the standard model. Hence $U(1)_B$ is explicitly broken by the weak interactions and is not an exact symmetry. Only $SU(3)_C$ remains as an exact symmetry of the standard model even after quantization.

Exact symmetries are very special in physics. Their existence points to something truly fundamental. Approximate symmetries, on the other hand, are ‘accidental’. A good example is isospin: the symmetry is due to the fact that u and d quarks have almost the same mass on the typical QCD scale Λ_{QCD} . The values of the quark masses are determined from Yukawa couplings to the Higgs field, which have nothing to do with Λ_{QCD} . Only by accident the u and d quark masses are small compared to it. Exact symmetries, on the other hand, should not be accidental. Instead they should be due to a dynamical gauge principle.

Now we want to derive the QCD Lagrange density from the gauge principle for the exact $SU(3)_C$ symmetry. For this purpose we consider transformations

$$g(x) \in SU(3)_C, \quad g^+(x)g(x) = g(x)g^+(x) = 1, \quad \det g(x) = 1, \quad (6.4.13)$$

and we now transform the quark spinors locally, i.e. we perform a gauge transformation

$$\Psi'_f(x) = g(x)\Psi_f(x). \quad (6.4.14)$$

The Lagrange density of the gauge transformed fields is then given by

$$\begin{aligned} \mathcal{L}(\bar{\Psi}', \Psi') &= \sum_f \bar{\Psi}_f(x)g^+(x)(i\gamma^\mu\partial_\mu - m_f)g(x)\Psi_f(x) \\ &= \sum_f \bar{\Psi}_f(x)(i\gamma^\mu\partial_\mu + i\gamma^\mu g^+(x)\partial_\mu g(x) - m_f)\Psi_f(x). \end{aligned} \quad (6.4.15)$$

Due to the term $i\gamma^\mu g^+(x)\partial_\mu g(x)$ the Lagrange density is not gauge invariant. To investigate this term further we write

$$g(x) = \exp(iH(x)), \quad (6.4.16)$$

where $H(x) \in su(3)_C$ is an element of the algebra, i.e.

$$H^+(x) = H(x), \quad \text{Tr}H(x) = 0. \quad (6.4.17)$$

It is instructive to convince oneself that $ig^+(x)\partial_\mu g(x)$ is an element of the algebra. Next we introduce an algebra-valued gauge potential

$$G_\mu(x) = ig_s G_\mu^a(x)\lambda_a, \quad G_\mu^a(x) \in R, \quad a \in \{1, 2, \dots, 8\}, \quad (6.4.18)$$

whose gauge variation is supposed to cancel $ig^+(x)\partial_\mu g(x)$. Here g_s is the dimensionless gauge coupling constant of the strong interactions. We postulate the usual behavior under gauge transformations

$$G'_\mu(x) = g(x)(G_\mu(x) + \partial_\mu)g^+(x), \quad (6.4.19)$$

because then the modified Lagrange function

$$\mathcal{L}(\bar{\Psi}, \Psi, G_\mu) = \sum_f \bar{\Psi}_f(x)(i\gamma^\mu(G_\mu + \partial_\mu) - m_f)\Psi_f(x) \quad (6.4.20)$$

is gauge invariant. It is a good exercise to show this explicitly.

The kinetic for the gluons term is still missing in the QCD Lagrange density. The gluon field strength is

$$G_{\mu\nu}(x) = \partial_\mu G_\nu(x) - \partial_\nu G_\mu(x) + [G_\mu(x), G_\nu(x)]. \quad (6.4.21)$$

In contrast to an Abelian gauge theory the field strength is not gauge invariant. It transforms as

$$G'_{\mu\nu}(x) = g(x)G_{\mu\nu}(x)g^+(x). \quad (6.4.22)$$

The full Lagrange function of QCD finally takes the form

$$\begin{aligned} \mathcal{L}_{QCD}(\bar{\Psi}, \Psi, G_\mu) &= \sum_f \bar{\Psi}_f(x)(i\gamma^\mu(G_\mu + \partial_\mu) - m_f)\Psi_f(x) \\ &- \frac{1}{2g_s^2} \text{Tr} G^{\mu\nu}(x)G_{\mu\nu}(x). \end{aligned} \quad (6.4.23)$$

It is instructive to show that $\mathcal{L}_{QCD}(\bar{\Psi}, \Psi, G_\mu)$ is indeed gauge invariant.

Using gauge invariance as the guiding principle one can construct another term in the QCD Lagrange function

$$\mathcal{L}_{QCD}^{\theta}(\bar{\Psi}, \Psi, G_{\mu}) = \mathcal{L}_{QCD}(\bar{\Psi}, \Psi, G_{\mu}) + \frac{\theta}{32\pi^2} \varepsilon^{\mu\nu\rho\sigma} \text{Tr} G_{\mu\nu}(x) G_{\rho\sigma}. \quad (6.4.24)$$

The role of the prefactor $32\pi^2$ will become clear when we discuss the $U(1)$ -problem. The parameter θ is the vacuum angle of QCD. For $\theta \neq 0$ the CP symmetry would be explicitly broken in the strong interactions (as it is in fact the case for the weak interactions). This would lead to a nonvanishing electric dipole moment of the neutron. The corresponding experimental result is compatible with zero, and one concludes that

$$|\theta| < 10^{-9}. \quad (6.4.25)$$

The question arises why θ is compatible with zero. This is the so-called strong CP problem, which is still unsolved, although various explanations have been suggested. The solution of the strong CP problem goes most probably beyond QCD, and perhaps beyond the Standard model. Therefore it is of no concern for us in this course.

If one wants to construct a renormalizable QCD theory in four dimensions, one cannot add any further terms to the Lagrangian. This follows from gauge invariance and the dimensions of quark and gluon fields.

An important difference between Abelian and non-Abelian gauge theories is that in a non-Abelian gauge theory the gauge fields are themselves charged. In fact, they transform in the adjoint representation of the gauge group — the $SU(3)_C$ octet in case of QCD. Correspondingly, there are eight gluons $G_{\mu}^a(x)$, $a \in \{1, 2, \dots, 8\}$. The non-Abelian charge of the gluons leads to a self interaction, that is not present for the Abelian photons. The interaction results from the commutator term in the gluon field strength. It gives rise to three and four gluon vertices in the QCD Feynman rules. We will not derive the QCD Feynman rules here, we discuss them only qualitatively. The terms in the Lagrange function, that are quadratic in $G_{\mu}(x)$ give rise to the free gluon propagator. Due to the commutator term, however, there are also terms cubic and quartic in $G_{\mu}(x)$, that lead to the gluon self interaction vertices. Correspondingly, there is a free quark propagator and a quark-gluon vertex. The perturbative quantization of a non-Abelian gauge theory requires to fix the gauge. In the Landau gauge $\partial^{\mu} G_{\mu}(x) = 0$ this leads to so-called ghost fields, which are scalars, but still anticommute. Correspondingly, there is a ghost propagator and a ghost-gluon vertex. In QCD the ghost fields are also color octets. They are only a mathematical tool arising in the loops of a Feynman diagram, not in external legs. Strictly speaking one

could say the same about quarks and gluons, because they also cannot exist as asymptotic states.

The objects in the classical QCD Lagrange function do not directly correspond to observable quantities. Both fields and coupling constants get renormalized. In particular, the formal expression

$$Z = \int \mathcal{D}\bar{\Psi}\mathcal{D}\Psi\mathcal{D}G \exp(-i \int d^4x \mathcal{L}_{QCD}(\bar{\Psi}, \Psi, G_\mu)) \quad (6.4.26)$$

for the QCD path integral is undefined, i.e. divergent, until it is regularized, i.e. made finite, and appropriately renormalized. In gauge theories it is essential that gauge invariance is maintained in the regularized theory. A regularization scheme that allows nonperturbative calculations defines the path integral on a space-time lattice with spacing ε . The renormalization of the theory corresponds to performing the continuum limit $\varepsilon \rightarrow 0$ in a controlled way, such that ratios of particle masses — i.e. the physics — remains constant. A perturbative regularization scheme works with single Feynman diagrams. The loop integrations in the corresponding mathematical expressions can be divergent in four dimensions. In dimensional regularization one works in d dimensions (by analytic continuation in d) and one performs the limit $\varepsilon = 4 - d \rightarrow 0$ again such that the physics remains constant. To absorb the divergencies quarks and gluon fields are renormalized

$$\Psi(x) = Z_\Psi(\varepsilon)^{1/2} \Psi^r(x), \quad G_\mu(x) = Z_G(\varepsilon)^{1/2} G_\mu^r(x), \quad (6.4.27)$$

and also the coupling constant is renormalized via

$$g_s = \frac{Z(\varepsilon)}{Z_\Psi(\varepsilon)Z_G(\varepsilon)^{1/2}} g_s^r. \quad (6.4.28)$$

Here the unrenormalized quantities as well as the Z -factors are divergent, but the renormalized quantities are finite in the limit $\varepsilon \rightarrow 0$. Correspondingly, one renormalizes the n -point Green's functions and the resulting vertex functions

$$\Gamma_{n_\Psi, n_G}^r(k_i, p_j) = \lim_{\varepsilon \rightarrow 0} Z_\Psi(\varepsilon)^{n_\Psi/2} Z_G(\varepsilon)^{n_G/2} \Gamma_{n_\Psi, n_G}(k_i, p_j, \varepsilon). \quad (6.4.29)$$

Demanding convergence of the renormalized vertex function fixes the divergent part of the Z -factors. To fix the finite part as well one must specify so-called renormalization conditions. In QCD this can be done using the vertex functions $\Gamma_{0,2}$, $\Gamma_{2,0}$ and $\Gamma_{2,1}$, i.e. the inverse gluon and quark propagators and the quark-gluon vertex. As opposed to QED, where mass and charge of the electron are directly observable, in QCD one chooses an arbitrary scale \mathcal{M} to formulate the

renormalization conditions

$$\begin{aligned}\Gamma_{0,2}^r(p, -p)_{ab}^{\mu\nu}|_{p^2=-\mathcal{M}^2} &= i(-g_{\mu\nu}p^2 + p^\mu p^\nu)\delta_{ab}, \\ \Gamma_{2,0}^r(k, k)|_{k^2=-\mathcal{M}^2} &= i\gamma^\mu k_\mu, \\ \Gamma_{2,1}^r(k, k, k)_a^\mu|_{k^2=-\mathcal{M}^2} &= -ig_s^r \frac{\lambda_a}{2} \gamma^\mu.\end{aligned}\tag{6.4.30}$$

The renormalized vertex functions are functions of the renormalized coupling constant g_s^r and of the renormalization scale \mathcal{M} , while the unrenormalized vertex functions depend on the bare coupling g_s and on the regularization parameter ε (the cut-off). Hence, there is a hidden relation

$$g_s^r = g_s^r(g, \varepsilon, \mathcal{M}).\tag{6.4.31}$$

This relation defines the so-called β -function

$$\beta(g_s^r) = \lim_{\varepsilon \rightarrow 0} \mathcal{M} \frac{\partial}{\partial \mathcal{M}} g_s^r(g, \varepsilon, \mathcal{M}).\tag{6.4.32}$$

The β -function can be computed in QCD perturbation theory. To leading order in the coupling constant one obtains

$$\beta(g_s^r) = -\frac{(g_s^r)^3}{16\pi^2} \left(11 - \frac{2}{3}N_f\right).\tag{6.4.33}$$

Here N_f is the number of quark flavors (massless quarks have been assumed). Fixed points g_s^* of the renormalization group are of special interest. They are invariant under a change of the arbitrarily chosen renormalization scale \mathcal{M} , and hence they correspond to zeros of the β -function. In QCD there is a single fixed point at $g_s^* = 0$. For

$$11 - \frac{2}{3}N_f > 0 \Rightarrow N_f \leq 16,\tag{6.4.34}$$

i.e. for not too many flavors, the β -function is negative close to the fixed point. This behavior is known as asymptotic freedom. It is typical for non-Abelian gauge theories in four dimensions, as long as there are not too many fermions or scalars. Asymptotic freedom is due to the self interaction of the gauge field, that is not present in an Abelian theory. We now use

$$\begin{aligned}\beta(g_s^r) &= \mathcal{M} \frac{\partial}{\partial \mathcal{M}} g_s^r = -\frac{(g_s^r)^3}{16\pi^2} \left(11 - \frac{2}{3}N_f\right) \Rightarrow \\ \frac{\partial g_s^r}{\partial \mathcal{M}} / (g_s^r)^3 &= \frac{1}{2} \frac{\partial (g_s^r)^2}{\partial \mathcal{M}} / (g_s^r)^4 = -\frac{11 - \frac{2}{3}N_f}{16\pi^2} \frac{1}{\mathcal{M}} \Rightarrow \\ \frac{\partial (g_s^r)^2}{(g_s^r)^4} &= -\frac{33 - 2N_f}{24\pi^2} \frac{\partial \mathcal{M}}{\mathcal{M}} \Rightarrow \frac{1}{(g_s^r)^2} = \frac{33 - 2N_f}{24\pi^2} \log \frac{\mathcal{M}}{\Lambda_{QCD}}.\end{aligned}\tag{6.4.35}$$

Here Λ_{QCD} is an integration constant. Introducing the renormalized strong fine structure constant

$$\alpha_s^r = \frac{(g_s^r)^2}{4\pi}, \quad (6.4.36)$$

we obtain

$$\alpha_s^r(\mathcal{M}) = \frac{6\pi}{33 - 2N_f} \frac{1}{\log(\mathcal{M}/\Lambda_{QCD})}. \quad (6.4.37)$$

At high energy scales \mathcal{M} the renormalized coupling constant slowly (i.e. logarithmically) goes to zero. Hence the quarks then behave as free particles. This confirms the central assumption of the parton model, such that we can finally identify the partons as quarks and gluons.

The classical Lagrange function for QCD with massless fermions has no dimensionful parameter. Hence the classical theory is scale invariant, i.e. to each solution with energy E correspond other solutions with scaled energy λE for any arbitrary scale parameter λ . Scale invariance, however, is anomalous. It does not survive the quantization of the theory. This explains why there is a proton with a very specific mass $E = M_p$, but no scaled version of it with mass λM_p . We now understand better why this is the case. In the process of quantization the dimensionful scale \mathcal{M} (and related to this Λ_{QCD}) emerged, leading to an explicit breaking of the scale invariance of the classical theory. Scale transformations are therefore no symmetry of QCD.

6.5 Spontaneous Chiral Symmetry Breaking

Chiral symmetry is an approximate global symmetry of the QCD Lagrange density, that results from the fact that the u and d quark masses are small compared to the typical QCD scale Λ_{QCD} . Neglecting the quark masses the QCD Lagrange density is invariant against separate $U(2)$ transformations of the left and right handed quarks, such that we have a $U(2)_L \otimes U(2)_R$ symmetry. We can decompose each $U(2)$ symmetry into an $SU(2)$ and a $U(1)$ part, and hence we obtain $SU(2)_L \otimes SU(2)_R \otimes U(1)_L \otimes U(1)_R$. The $U(1)_B$ symmetry related to baryon number conservation corresponds to simultaneous rotations of left and right handed quarks, i.e. $U(1)_B = U(1)_{L=R}$. The remaining so-called axial $U(1)$ is affected by the Adler-Bell-Jackiw anomaly. It is explicitly broken by quantum effects, and hence it is not a symmetry of QCD. In the next chapter we return to the $U(1)$ problem related to this symmetry. Here we are interested in the ordinary (non anomalous) symmetries of QCD — the $SU(2)_L \otimes SU(2)_R \otimes U(1)_B$ chiral symmetry. Based on this symmetry one would expect corresponding degeneracies in the QCD spectrum. Indeed we saw that the hadrons can be classified

as isospin multiplets. The isospin transformations are $SU(2)_I$ rotations, that act on left and right handed fermions simultaneously, i.e. $SU(2)_I = SU(2)_{L=R}$. The symmetry that is manifest in the spectrum is hence $SU(2)_I \otimes U(1)_B$, but not the full chiral symmetry $SU(2)_L \otimes SU(2)_R \otimes U(1)_B$. One concludes that chiral symmetry must be spontaneously broken. The order parameter of chiral symmetry breaking is the so-called chiral condensate $\langle \bar{\Psi}\Psi \rangle$. When a continuous global symmetry breaks spontaneously, massless particles — the so-called Goldstone bosons — appear in the spectrum. According to the Goldstone theorem the number of Goldstone bosons is the difference of the number of generators of the full symmetry group and the subgroup remaining after spontaneous breaking. In our case we hence expect $3 + 3 + 1 - 3 - 1 = 3$ Goldstone bosons. In QCD they are identified as the pions π^+ , π^0 and π^- . Of course, the pions are light, but they are not massless. This is due to a small explicit chiral symmetry breaking related to the small but nonzero masses of the u and d quarks. Chiral symmetry is only an approximate symmetry, and the pions are only pseudo Goldstone bosons. It turns out that the pion mass squared is proportional to the quark mass. When we also consider the s quark as being light, chiral symmetry can be extended to $SU(3)_L \otimes SU(3)_R \otimes U(1)_B$, which then breaks spontaneously to $SU(3)_F \otimes U(1)_B$. Then one expects $8 + 8 + 1 - 8 - 1 = 8$ Goldstone bosons. The five additional bosons are identified as the four kaons K^+ , K^0 , \bar{K}^0 , K^- and the η -meson. Since the s quark mass is not really negligible, these pseudo Goldstone bosons are heavier than the pion.

The Goldstone bosons are the lightest particles in QCD. Therefore they determine the dynamics at small energies. One can construct effective theories that are applicable in the low energy regime, and that are formulated in terms of Goldstone boson fields. At low energies the Goldstone bosons interact only weakly, and can hence be treated perturbatively. This is done systematically in so-called chiral perturbation theory.

Let us consider the quark part of the QCD Lagrange density

$$\mathcal{L}(\bar{\Psi}, \Psi, G_\mu) = \bar{\Psi}(x)(i\gamma^\mu(G_\mu(x) + \partial_\mu) - \mathcal{M})\Psi(x). \quad (6.5.1)$$

We now decompose the quark fields in right and left handed components

$$\Psi_R(x) = \frac{1}{2}(1+\gamma_5)\Psi(x), \quad \Psi_L(x) = \frac{1}{2}(1-\gamma_5)\Psi(x), \quad \Psi(x) = \Psi_R(x) + \Psi_L(x). \quad (6.5.2)$$

Here we have used

$$\gamma_5 = i\gamma^0\gamma^1\gamma^2\gamma^3, \quad \{\gamma^\mu, \gamma^\nu\} = 2g_{\mu\nu}, \quad \{\gamma^\mu, \gamma_5\} = 0. \quad (6.5.3)$$

Next we consider the adjoint spinors

$$\begin{aligned}
\bar{\Psi}_R(x) &= \Psi_R(x)^\dagger \gamma^0 = \Psi(x)^\dagger \frac{1}{2}(1 + \gamma_5^+) \gamma^0 = \Psi(x)^\dagger \gamma^0 \frac{1}{2}(1 - \gamma_5) \\
&= \bar{\Psi}(x) \frac{1}{2}(1 - \gamma_5), \\
\bar{\Psi}_L(x) &= \Psi_L(x)^\dagger \gamma^0 = \Psi(x)^\dagger \frac{1}{2}(1 - \gamma_5^+) \gamma^0 = \Psi(x)^\dagger \gamma^0 \frac{1}{2}(1 + \gamma_5) \\
&= \bar{\Psi}(x) \frac{1}{2}(1 + \gamma_5).
\end{aligned} \tag{6.5.4}$$

Here we used

$$\gamma^0 \gamma_5^+ \gamma^0 = -\gamma_5. \tag{6.5.5}$$

Inserting the decomposed spinors in the Lagrange density we obtain

$$\mathcal{L}(\bar{\Psi}, \Psi, G_\mu) = (\bar{\Psi}_R(x) + \bar{\Psi}_L(x))(i\gamma^\mu(G_\mu(x) + \partial_\mu) - \mathcal{M})(\Psi_R(x) + \Psi_L(x)). \tag{6.5.6}$$

First, we investigate the γ^μ term

$$\begin{aligned}
&\bar{\Psi}_R(x) i\gamma^\mu(G_\mu(x) + \partial_\mu) \Psi_L(x) \\
&= \bar{\Psi}(x) \frac{1}{2}(1 - \gamma_5) i\gamma^\mu(G_\mu(x) + \partial_\mu) \frac{1}{2}(1 - \gamma_5) \Psi(x) \\
&= \bar{\Psi}(x) \frac{1}{4}(1 - \gamma_5)(1 + \gamma_5) i\gamma^\mu(G_\mu(x) + \partial_\mu) \Psi(x) = 0.
\end{aligned} \tag{6.5.7}$$

On the other hand, for the mass term one finds

$$\begin{aligned}
\bar{\Psi}_R(x) \mathcal{M} \Psi_R(x) &= \bar{\Psi}(x) \frac{1}{2}(1 - \gamma_5) \mathcal{M} \frac{1}{2}(1 + \gamma_5) \Psi(x) \\
&= \bar{\Psi}(x) \frac{1}{4}(1 - \gamma_5)(1 + \gamma_5) \mathcal{M} \Psi(x) = 0.
\end{aligned} \tag{6.5.8}$$

Hence, we can write

$$\begin{aligned}
\mathcal{L}(\bar{\Psi}, \Psi, G_\mu) &= \bar{\Psi}_R(x) i\gamma^\mu(G_\mu(x) + \partial_\mu) \Psi_R(x) \\
&+ \bar{\Psi}_L(x) i\gamma^\mu(G_\mu(x) + \partial_\mu) \Psi_L(x) \\
&- \bar{\Psi}_R(x) \mathcal{M} \Psi_L(x) - \bar{\Psi}_L(x) \mathcal{M} \Psi_R(x).
\end{aligned} \tag{6.5.9}$$

The γ^μ term decomposes into two decoupled contributions from right and left handed quarks. This part of the Lagrange density is invariant against separate $U(N_f)$ transformations of the right and left handed components in flavor space

$$\begin{aligned}
\Psi'_R(x) &= R \Psi_R(x), \quad \bar{\Psi}'(x) = \bar{\Psi}_R(x) R^+, \quad R \in U(N_f)_R, \\
\Psi'_L(x) &= L \Psi_L(x), \quad \bar{\Psi}'(x) = \bar{\Psi}_L(x) L^+, \quad L \in U(N_f)_L.
\end{aligned} \tag{6.5.10}$$

Without the mass term the classical QCD Lagrange density hence has a $U(N_f)_L \otimes U(N_f)_R$ symmetry. Due to the anomaly in the axial $U(1)$ symmetry the symmetry of the quantum theory is reduced to

$$SU(N_f)_L \otimes SU(N_f)_R \otimes U(1)_{L=R} = SU(N_f)_L \otimes SU(N_f)_R \otimes U(1)_B. \quad (6.5.11)$$

Of course, the chiral symmetry is only approximate, because the mass term couples right and left handed fermions. In addition, the mass matrix does not commute with R and L . If all quarks had the same mass, i.e. if $\mathcal{M} = m1$, one would have

$$\bar{\Psi}'_R(x) \mathcal{M} \Psi'_L(x) = \bar{\Psi}_R(x) R^+ m 1 L \Psi_L(x) = \bar{\Psi}_R(x) R^+ L \mathcal{M} \Psi_L(x). \quad (6.5.12)$$

Then the mass term is invariant only against simultaneous transformations $R = L$ such that $R^+ L = R^+ R = 1$. Hence, chiral symmetry is then explicitly broken to

$$SU(N_f)_{L=R} \otimes U(1)_{L=R} = SU(N_f)_F \otimes U(1)_B, \quad (6.5.13)$$

which corresponds to the flavor and baryon number symmetry. In reality the quark masses are different, and the symmetry is in fact explicitly broken to

$$\otimes_f U(1)_f = U(1)_u \otimes U(1)_d \otimes U(1)_s. \quad (6.5.14)$$

It is, however, much more important that the u and d quark masses are small, and can hence almost be neglected. Therefore, in reality the chiral $SU(2)_L \otimes SU(2)_R \otimes U(1)_B \otimes U(1)_s$ symmetry is almost unbroken explicitly. Since the s quark is heavier, $SU(3)_L \otimes SU(3)_R \otimes U(1)_B$ is a more approximate chiral symmetry, because it is explicitly more strongly broken.

Since the masses of the u and d quarks are so small the $SU(2)_L \otimes SU(2)_R$ chiral symmetry should work very well. Hence, one would expect that the hadron spectrum shows corresponding degeneracies. Let us neglect quark masses and consider the then conserved currents

$$\begin{aligned} J_\mu^{La}(x) &= \bar{\Psi}_L(x) \gamma_\mu \frac{\sigma^a}{2} \Psi_L(x), \\ J_\mu^{Ra}(x) &= \bar{\Psi}_R(x) \gamma_\mu \frac{\sigma^a}{2} \Psi_R(x), \end{aligned} \quad (6.5.15)$$

where $a \in \{1, 2, 3\}$. From the right and left handed currents we now construct

vector and axial currents

$$\begin{aligned}
V_\mu^a(x) &= J_\mu^{La}(x) + J_\mu^{Ra}(x) \\
&= \bar{\Psi}(x) \frac{1}{2} (1 + \gamma_5) \gamma_\mu \frac{\sigma^a}{2} \frac{1}{2} (1 - \gamma_5) \Psi(x) \\
&+ \bar{\Psi}(x) \frac{1}{2} (1 - \gamma_5) \gamma_\mu \frac{\sigma^a}{2} \frac{1}{2} (1 + \gamma_5) \Psi(x) \\
&= \bar{\Psi}(x) \frac{1}{2} (1 + \gamma_5) \gamma_\mu \frac{\sigma^a}{2} \Psi(x) + \bar{\Psi}(x) \frac{1}{2} (1 - \gamma_5) \gamma_\mu \frac{\sigma^a}{2} \Psi(x) \\
&= \bar{\Psi}(x) \gamma_\mu \frac{\sigma^a}{2} \Psi(x), \\
A_\mu^a(x) &= J_\mu^{La}(x) - J_\mu^{Ra}(x) = \bar{\Psi}(x) \gamma_5 \gamma_\mu \frac{\sigma^a}{2} \Psi(x).
\end{aligned} \tag{6.5.16}$$

Let us consider an $SU(2)_L \otimes SU(2)_R$ invariant state $|\Phi\rangle$ as a candidate for the QCD vacuum. Then

$$\langle \Phi | J_\mu^{La}(x) J_\nu^{Rb}(y) | \Phi \rangle = \langle \Phi | J_\mu^{Ra}(x) J_\nu^{Lb}(y) | \Phi \rangle = 0, \tag{6.5.17}$$

and hence

$$\langle \Phi | V_\mu^a(x) V_\nu^b(y) | \Phi \rangle = \langle \Phi | A_\mu^a(x) A_\nu^b(y) | \Phi \rangle. \tag{6.5.18}$$

On both sides of the equation one can insert complete sets of states between the two operators. On the left hand side states with quantum numbers $J^P = 0^+, 1^-$ contribute, while on the right hand side the nonzero contributions come from states $0^-, 1^+$. The two expressions can be equal only if the corresponding parity partners are energetically degenerate. In the observed hadron spectrum there is no degeneracy of particles with even and odd parity, not even approximately. We conclude that the $SU(2)_L \otimes SU(2)_R$ invariant state $|\Phi\rangle$ is not the real QCD vacuum. The true vacuum $|0\rangle$ cannot be chirally invariant. The same is true for all other eigenstates of the QCD Hamiltonian. This means that chiral symmetry must be spontaneously broken.

Let us now consider the states

$$\begin{aligned}
Q_V^a |0\rangle &= \int d^3x V_0^a(\vec{x}, 0) |0\rangle, \\
Q_A^a |0\rangle &= \int d^3x A_0^a(\vec{x}, 0) |0\rangle,
\end{aligned} \tag{6.5.19}$$

constructed from the vacuum by acting with the vector and axial charge densities. If the vacuum were chirally symmetric we would have

$$Q_V^a |\Phi\rangle = Q_A^a |\Phi\rangle = 0. \tag{6.5.20}$$

The real QCD vacuum is not chirally invariant because

$$Q_A^a|0\rangle \neq 0. \quad (6.5.21)$$

Since the axial current is conserved (for massless quarks $\partial^\mu A_\mu^a(x) = 0$) we have

$$[H_{QCD}, Q_A^a] = 0. \quad (6.5.22)$$

Hence the new state $Q_A^a|0\rangle$ is again an eigenstate of the QCD Hamilton operator

$$H_{QCD}Q_A^a|0\rangle = Q_A^a H_{QCD}|0\rangle = 0 \quad (6.5.23)$$

with zero energy. This state corresponds to a massless Goldstone boson with quantum numbers $J^P = 0^-$. These pseudoscalar particles are identified with the pions of QCD.

If one would also have $Q_V^a|0\rangle \neq 0$, the vector flavor symmetry would also be spontaneously broken, and there would be another set of scalar Goldstone bosons with $J^P = 0^+$. Such particles do not exist in the hadron spectrum, and we conclude that the isospin symmetry $SU(2)_I = SU(2)_{L=R}$ is not spontaneously broken. As we have seen before, the isospin symmetry is indeed manifest in the hadron spectrum.

One can also detect spontaneous chiral symmetry breaking by investigating the chiral order parameter

$$\langle \bar{\Psi}\Psi \rangle = \langle 0|\bar{\Psi}(x)\Psi(x)|0\rangle = \langle 0|\bar{\Psi}_R(x)\Psi_L(x) + \bar{\Psi}_L(x)\Psi_R(x)|0\rangle. \quad (6.5.24)$$

The order parameter is invariant against simultaneous transformations $R = L$, but not against general chiral rotations. If chiral symmetry would be intact the chiral condensate would vanish. When the symmetry is spontaneously broken, on the other hand, $\langle \bar{\Psi}\Psi \rangle \neq 0$.

Being almost massless the Goldstone bosons are the lightest particles in QCD. Therefore they dominate the dynamics of the strong interactions at low energies, and it is possible to switch to a low energy effective description that only involves the Goldstone bosons. This is not only the case for QCD but also for any other model with a continuous global symmetry G breaking spontaneously to a subgroup H , provided that there are no other massless particles besides the Goldstone bosons. The Goldstone bosons are described by fields in the coset space G/H , in which points are identified if they are connected by symmetry transformations of the remaining subgroup H . In QCD we have

$$\begin{aligned} G &= SU(N_f)_L \otimes SU(N_f)_R \otimes U(1)_B, \\ H &= SU(N_f)_{L=R} \otimes U(1)_B, \end{aligned} \quad (6.5.25)$$

and the corresponding coset space is

$$G/H = SU(N_f). \quad (6.5.26)$$

Hence the Goldstone boson fields can be represented as special unitary matrices $U(x) \in SU(N_f)$. Under chiral rotations they transform as

$$U(x)' = LU(x)R^+. \quad (6.5.27)$$

Now we must construct an effective Lagrange function which is chirally symmetric. For this purpose we consider

$$\partial_\mu U(x)' = L\partial_\mu U(x)R^+, \quad (6.5.28)$$

and we form a chirally invariant Lorentz scalar

$$\mathcal{L}(U) = \frac{f_\pi^2}{4} \text{Tr}(\partial^\mu U(x)^+ \partial_\mu U(x)). \quad (6.5.29)$$

The coupling constant f_π determines the strength of the interaction between the Goldstone bosons. It also plays a role in the weak decay of the pion and is therefore known as the pion decay constant. The above Lagrange density is chiral invariant because

$$\begin{aligned} \mathcal{L}(U') &= \frac{f_\pi^2}{4} \text{Tr}(\partial^\mu U(x)'^+ \partial_\mu U(x)') \\ &= \frac{f_\pi^2}{4} \text{Tr}(R\partial^\mu U(x)^+ L^+ L\partial_\mu U(x)R^+) = \mathcal{L}(U). \end{aligned} \quad (6.5.30)$$

We still must introduce the chiral symmetry breaking mass terms, which enter via the quark mass matrix. We write

$$\mathcal{L}(U) = \frac{f_\pi^2}{4} \text{Tr}(\partial^\mu U(x)^+ \partial_\mu U(x)) + c \text{Tr}(\mathcal{M}(U + U^+)). \quad (6.5.31)$$

Under chiral transformations the additional term transforms into

$$\text{Tr}(\mathcal{M}(U(x)' + U(x)'^+)) = \text{Tr}(\mathcal{M}(LU(x)R^+ + RU(x)^+L^+)). \quad (6.5.32)$$

If all quark masses are equal, i.e. for $\mathcal{M} = m1$, the Lagrange density is again invariant against rotations $R^+L = 1$ and hence for $R = L$. For a general mass matrix the symmetry is reduced to $\otimes_f U(1)_f/U(1)_B$.

We still have to determine the prefactor c . For this purpose we determine the vacuum value of $\partial/\partial m_f \mathcal{L}|_{\mathcal{M}=0}$ first in QCD

$$\langle 0 | \frac{\partial}{\partial m_f} \mathcal{L}|_{\mathcal{M}=0} | 0 \rangle = \langle 0 | -\bar{\Psi}_f \Psi_f | 0 \rangle = -\frac{1}{N_f} \langle \bar{\Psi} \Psi \rangle. \quad (6.5.33)$$

In the effective theory the same value should arise. The classical vacuum of the effective theory with $\mathcal{M} = 0$ corresponds to a constant field

$$U(x) = 1. \quad (6.5.34)$$

Hence, in the effective theory we have

$$\langle 0 | \frac{\partial}{\partial m_f} \mathcal{L} |_{\mathcal{M}=0} | 0 \rangle = c \text{Tr}(\text{diag}(1, 0, \dots, 0)(1 + 1)) = 2c, \quad (6.5.35)$$

and therefore

$$c = -\frac{1}{2N_f} \langle \bar{\Psi} \Psi \rangle. \quad (6.5.36)$$

The constants f_π and $\langle \bar{\Psi} \Psi \rangle$ determine the low energy dynamics of QCD. They can only be determined from QCD itself, and must be inserted in the effective theory as a priori unknown parameters. Up to these two low energy constants the Goldstone boson dynamics is completely determined by chiral symmetry. At higher energies additional terms arise in the effective theory. Again, they are restricted by chiral symmetry requirements, and they contain new parameters. In fact, chiral perturbation theory is a systematic low energy expansion, in which the higher order terms contain a larger number of derivatives. Here we restrict ourselves to lowest order, and hence to the Lagrange density

$$\mathcal{L}(U) = \frac{f_\pi^2}{4} \text{Tr}(\partial^\mu U(x)^+ \partial_\mu U(x)) - \frac{1}{2N_f} \langle \bar{\Psi} \Psi \rangle \text{Tr}(\mathcal{M}(U + U^+)). \quad (6.5.37)$$

Chiral perturbation theory is an expansion around the classical vacuum solution $U(x) = 1$. One writes

$$U(x) = \exp(i\pi^a(x)\eta_a/f_\pi), \quad a \in \{1, 2, \dots, N_f^2 - 1\}, \quad (6.5.38)$$

where η_a are the generators of $SU(N_f)$, and one expands in powers of $\pi^a(x)$. To lowest order we have

$$U(x) = 1 + i\pi^a(x)\eta_a/f_\pi, \quad \partial_\mu U(x) = i\partial_\mu \pi^a(x)\eta_a/f_\pi, \quad (6.5.39)$$

and hence

$$\begin{aligned} \mathcal{L}(U) &= \frac{1}{4} \text{Tr}(\partial^\mu \pi^a(x)\eta_a \partial_\mu \pi^b(x)\eta_b) - \frac{1}{2N_f} \langle \bar{\Psi} \Psi \rangle \text{Tr}(\mathcal{M}(1 + 1)) \\ &= \frac{1}{2} \partial^\mu \pi^a(x) \partial_\mu \pi^b(x) \delta_{ab} - \frac{1}{N_f} \langle \bar{\Psi} \Psi \rangle \text{Tr} \mathcal{M}. \end{aligned} \quad (6.5.40)$$

The last term is an irrelevant constant. We have to expand consequently to order π^2

$$U(x) = 1 + i\pi^a(x)\eta_a/f_\pi + \frac{1}{2} (i\pi^a(x)\eta_a/f_\pi)^2, \quad (6.5.41)$$

such that for quarks with equal masses, i.e. for $\mathcal{M} = m1$,

$$\begin{aligned} \text{Tr}(\mathcal{M}(U(x) + U(x)^+)) &= m \text{Tr}(U(x) + U(x)^+) \\ &= 2mN_f - m \frac{1}{f_\pi^2} \pi^a(x) \pi^b(x) \text{Tr}(\eta_a \eta_b) \\ &= 2mN_f - 2m \frac{1}{f_\pi^2} \pi^a(x) \pi^a(x). \end{aligned} \quad (6.5.42)$$

Altogether we obtain

$$\mathcal{L}(U) = \frac{1}{2} \partial^\mu \pi^a(x) \partial_\mu \pi^a(x) - \frac{1}{N_f} \langle \bar{\Psi} \Psi \rangle (mN_f - m \frac{1}{f_\pi^2} \pi^a(x) \pi^a(x)). \quad (6.5.43)$$

The resulting equation of motion is given by

$$\partial^\mu \frac{\delta \mathcal{L}}{\delta \partial^\mu \pi^a} - \frac{\delta \mathcal{L}}{\delta \pi^a} = \partial^\mu \partial_\mu \pi^a(x) - \frac{2m \langle \bar{\Psi} \Psi \rangle}{N_f f_\pi^2} \pi^a(x) = 0. \quad (6.5.44)$$

This is the Klein-Gordon equation for a pseudoscalar particle with mass

$$M_\pi^2 = \frac{2m \langle \bar{\Psi} \Psi \rangle}{N_f f_\pi^2}. \quad (6.5.45)$$

This behavior is typical for the mass of a pseudo Goldstone boson: it is proportional to the square root of the explicit symmetry breaking term. It is very instructive to derive a corresponding mass formula for a nondegenerate mass matrix, e.g. for light quark masses $m_u = m_d = m_q$, but a heavier strange quark mass m_s . For general quark masses one obtains

$$\begin{aligned} M_{K^+}^2 &= M_{K^-}^2 = \frac{(m_u + m_s) \langle \bar{\Psi} \Psi \rangle}{N_f f_\pi^2}, \\ M_{K^0}^2 &= M_{\bar{K}^0}^2 = \frac{(m_d + m_s) \langle \bar{\Psi} \Psi \rangle}{N_f f_\pi^2}, \\ M_\eta^2 &= \frac{1}{3} \frac{(m_u + m_d + 4m_s) \langle \bar{\Psi} \Psi \rangle}{N_f f_\pi^2}. \end{aligned} \quad (6.5.46)$$

This leads to the following mass relation

$$3M_\eta^2 + M_\pi^2 = 2M_{K^+}^2 + 2M_{K^0}^2, \quad (6.5.47)$$

which is well satisfied experimentally. The left hand side has a value of 0.923GeV^2 and the right hand side is 0.984GeV^2 . Introducing the average light quark mass

$$m_q = \frac{1}{2}(m_u + m_d) \quad (6.5.48)$$

one obtains

$$\frac{M_{K^+}^2 + M_{K^0}^2}{M_\pi^2} = \frac{m_s + m_q}{m_q}, \quad (6.5.49)$$

which yields $m_s/m_q = 24.2$. Similarly

$$\frac{M_\eta^2}{M_\pi^2} = \frac{2m_s + m_q}{3m_q}, \quad (6.5.50)$$

which leads to $m_s/m_q = 22.7$. Still the quark masses themselves are not directly measurable. Using various models one obtains

$$m_q \approx 0.007\text{GeV}, \quad (6.5.51)$$

and hence

$$m_s \approx 0.16\text{GeV}. \quad (6.5.52)$$

These are so-called current quark masses. They are much smaller than the constituent quark masses. Still, the mass difference is consistent with the value obtained in the constituent quark model.

Chapter 7

Electroweak and QCD Phase Transitions

Symmetries that are spontaneously broken at low temperatures, are often restored at high temperatures. Known examples are the melting of crystal lattices (restoration of translation invariance) or the transition from a superconductor to an ordinary conductor (restoration of $U(1)$ gauge invariance). As we have seen, also in the standard model of particle physics the ground state has several broken symmetries. First, the $SU(2)_L \otimes U(1)_Y$ gauge invariance is spontaneously broken to $U(1)_{em}$ — the gauge group of electromagnetism. The scale of the corresponding interactions is set by the vacuum value v of the scalar field, which is in the 100 GeV range. One would expect that the $SU(2)_L \otimes U(1)_Y$ symmetry gets restored in that temperature range. One may say that the condensate of the scalar field melts at very high temperatures. In the early Universe this effect occurs at about 10^{-12} sec after the Big Bang. Of course, the phase transition proceeds from high to low temperatures. Hence, the scalar condensate “crystallizes” out of the high temperature phase.

Similarly, the chiral symmetry of QCD is spontaneously broken at low temperatures, and the order parameter is the chiral condensate $\langle \bar{\Psi}\Psi \rangle$. The typical scale for the breaking of chiral symmetry is Λ_{QCD} which is in the 100 MeV range. Correspondingly, one expects a QCD phase transition at about 10^{-5} sec after the Bang. In the high temperature phase chiral symmetry is intact, and quarks are not confined — i.e. one has a quark-gluon plasma. At lower temperatures confinement sets in, quarks and gluons are combined to hadrons, and chiral symmetry is spontaneously broken.

A phase transition in the early Universe can, in principle, have drastic consequences for the cosmic evolution. In particular, at a first order phase transition, which is characterized by discontinuities in thermodynamic quantities, the system gets out of thermal equilibrium. At a very strong first order phase transition the high temperature phase is super-cooled, and then suddenly turns into the low temperature phase via bubble nucleation. This leads to inhomogeneities, which could even affect the quality of the Robertson-Walker ansatz for the metric. Even if inhomogeneities are not that strong, they can lead to fluctuations in the baryon density, with potential consequences for primordial nucleosynthesis.

The question if a phase transition is first or second order is a complicated dynamical problem, that generally can not be solved analytically. In the case of second order phase transitions, however, one can use universality arguments and study simple representatives of a universality class. For example, instead of the Higgs sector of the standard model or instead of QCD with two light quarks, one can investigate the physics of magnets, and then translate the results back to particle physics.

In QCD without quarks ($SU(3)$ pure gauge theory) the deconfinement phase transition is first order, and hence not universal. At high temperatures the $\mathbb{Z}(3)$ center symmetry of the $SU(3)$ gauge group is spontaneously broken. The different $\mathbb{Z}(3)$ phases are separated by domain walls. These walls are completely wet by the confined phase when one approaches the phase transition. Complete wetting is a universal phenomenon of interfaces, that occurs for hot gluons, but also for the liquid of tears in the human eye.

7.1 The QCD Phase Transition

As we have seen in the previous chapter, quantum systems (in particular quantum field theories) at finite temperature characterized by a partition function

$$Z = \text{Tr} \exp(-\beta H), \quad (7.1.1)$$

can be represented by a Euclidean path integral with a Euclidean time extent given by the inverse temperature $\beta = 1/T$. Due to the trace one integrates over periodic field configurations. At high temperatures the system is very short in the Euclidean time direction, and it effectively reduces to a 3-d system in the infinite temperature limit. Even at finite temperature, if the system undergoes a second order phase transition, the finite size in Euclidean time is negligible compared to the diverging correlation length, such that again dimensional reduction occurs.

In QCD with N_f massless quarks the $SU(N_f)_L \otimes SU(N_f)_R$ chiral symmetry is spontaneously broken at low temperatures. At high temperatures, on the other hand, one expects that, due to asymptotic freedom, quarks and gluons behave like free particles, and hence the theory deconfines and chiral symmetry should be restored. Lattice simulations indeed support this expectation. In the early Universe the deconfinement-confinement phase transition must have happened at a temperature of about 0.15 GeV. For two massless quarks the chiral symmetry group is $SU(2)_L \otimes SU(2)_R$, which is isomorphic to $O(4)$. At low temperatures the symmetry breaks spontaneously to $SU(2)_{L=R}$ which is isomorphic to $O(3)$. From condensed matter physics investigations of ferro- and antiferromagnets it is well known that 3-d systems with symmetry breaking $O(4) \rightarrow O(3)$ have second order phase transitions. Hence, based on universality and the fact that at high temperature QCD is dimensionally reduced to a 3-d theory, one expects that with two massless quarks the QCD phase transition would be second order. In fact, QCD would behave very similar to a magnet, e.g. all critical exponents defined at the phase transition would be the same. For three massless quarks, on the other hand, one expects a first order phase transition, because systems with symmetry breaking $SU(3)_L \otimes SU(3)_R \rightarrow SU(3)_{L=R}$ do not have second order transitions. In that case there is no universal behavior. In reality the quarks are not completely massless. Especially the s-quark has an intermediate mass, and it is unclear if the two or three flavor argument applies to our Universe. In any case, a true second order phase transition would only happen in the limit of vanishing quark mass. Therefore, in our universe we either had a first order phase transition or merely a cross over.

A first order phase transition proceeds via bubble nucleation, just as in boiling water. At the phase transition, both phases coexist with each other, and they are separated by an interface with a finite interface tension. In case of a first order QCD phase transition bubbles of confined phase would appear from the hot quark-gluon plasma. The bubbles would expand and would ultimately fill the whole universe. Witten has suggested that a strange state of matter — so-called quark nuggets — may be created during the QCD phase transition provided it is strongly first order. He argued that the expanding bubble walls of confined phase may compress the remaining quark-gluon plasma to super-large hadrons, which would be stable due to a large s-quark content that would be favored by the Pauli principle. These quark nuggets would have masses of about a ton. They would not participate in primordial nucleosynthesis and would actually be exotic candidates for baryonic dark matter. Still, if such objects existed in the universe, one would expect them to collide with the earth sometimes causing major damage. The fact, that we do not experience catastrophes like that too often, puts an upper limit on the number of quark nuggets. In fact, they seem to

be a very exotic form of matter, for which we have no observational evidence.

7.2 Phase Transition in the Purge Glue Theory

Let us consider an $SU(3)$ purge gauge theory (without quarks) at finite temperature $T = 1/\beta$, i.e. with periodic boundary conditions in the Euclidean time direction

$$G_\mu(\vec{x}, t + \beta) = G_\mu(\vec{x}, t). \quad (7.2.1)$$

Then we can construct the Polyakov loop

$$\Phi(\vec{x}) = \text{Tr}(\mathcal{P} \exp \int_0^\beta dt G_4(\vec{x}, t)) \in \mathbb{C}, \quad (7.2.2)$$

which is a complex scalar field in three dimensions. Like the Wilson loop measures the energy of a static quark-antiquark pair, the Polyakov loop measures the free energy F of a single static quark, i.e.

$$\langle \Phi(\vec{x}) \rangle \propto \exp(-\beta F). \quad (7.2.3)$$

In the confined phase, i.e. at low temperatures, quarks are confined and hence a single quark has infinite energy $F = \infty$. Consequently, $\langle \Phi(\vec{x}) \rangle = 0$ in the confined phase. In the deconfined phase, on the other hand, one expects quarks to deconfine such that then $\langle \Phi(\vec{x}) \rangle \neq 0$. This means that the Polyakov loop is an order parameter for the deconfinement phase transition in the pure gluon theory (with external static quarks). The Polyakov loop is invariant under ordinary periodic gauge transformations, but it transforms into

$$\Phi'(\vec{x}) = \Phi(\vec{x})z^*, \quad (7.2.4)$$

under transformations

$$g(\vec{x}, t + \beta) = g(\vec{x}, t)z, \quad (7.2.5)$$

which are periodic only up to a center element $z \in \mathbb{Z}(3)$. Hence, $\langle \Phi(\vec{x}) \rangle \neq 0$ signals the spontaneous breakdown of the $\mathbb{Z}(3)$ center symmetry. It is interesting that the symmetry is then broken spontaneously at high temperatures, in contrast to typical condensed matter examples. The spontaneous breakdown of $\mathbb{Z}(3)$ has been observed in lattice simulations of the pure $SU(3)$ gauge theory, and it has been clearly established that the deconfinement phase transition is first order. For two colors (i.e. for $SU(2)$ gauge theory) again the $\mathbb{Z}(2)$ center symmetry gets spontaneously broken, but then the phase transition turns out to be second order.

7.3 Domain Walls and Gluonic Interfaces

When a discrete symmetry is spontaneously broken, different phases exist that are distinguished by some order parameter. Different regions of space may be in different phases, and the order parameter changes as one goes from one region to another. Since the change in the order parameter costs free energy, the different regions are separated by domain walls with a free energy proportional to their surface area. Domain walls arise, for example, in the Ising model, where a $\mathbb{Z}(2)$ symmetry gets spontaneously broken. In the standard model of particle physics no discrete symmetry is spontaneously broken and domain walls do not arise. However, when one neglects quarks — as it is done in the quenched approximation to QCD — the $\mathbb{Z}(3)$ center symmetry of the $SU(3)$ gauge group plays a crucial role. At high temperatures it is spontaneously broken and there are in fact three distinct deconfined phases that are characterized by the values of the Polyakov loop. When different regions of space are in different phases, they are separated by deconfined-deconfined domain walls. When the temperature decreases the gluonic system undergoes a phase transition to the confined glueball phase. The phase transition is of first order, such that the confined and the three deconfined phases coexist at the critical temperature. Hence, at T_c there exist in addition confined-deconfined interfaces.

As we will see, at T_c the deconfined-deconfined interface tension is two times the confined-deconfined interface tension. This leads to the phenomenon of complete wetting: a deconfined-deconfined domain wall splits into two confined-deconfined interfaces with a layer of confined phase in between. Complete wetting is a critical phenomenon of interfaces, which arises although the bulk phase transition is first order. The critical behavior is characterized by the divergence of the thickness of the confined complete wetting layer, which is governed by a certain critical exponent. We will derive the values of the critical exponents of complete wetting from an effective 3-dimensional $\mathbb{Z}(3)$ symmetric Φ^4 model for the Polyakov loop, which is in the same universality class as the pure $SU(3)$ gauge theory. The shape of the interfaces (and also the values of the critical exponents) follow from the classical equations of motion of the effective theory. Complete wetting is a peculiarity of the pure gauge theory. When quarks are included they break the $\mathbb{Z}(3)$ symmetry explicitly, and two of the three deconfined phases become unstable (or at least metastable). Then stable deconfined-deconfined interfaces no longer exist at the phase transition and wetting does not arise. In fact, a region of space that was originally in a metastable $\mathbb{Z}(3)$ phase of the early Universe, converts into the stable phase about 10^{-14} seconds after the Big Bang, while the deconfinement phase transition occurred at about 10^{-5} seconds after

the Bang.

In full QCD the dynamics of a first order phase transition is governed by bubble nucleation: the early high temperature phase super-cools until the latent heat is large enough to provide the necessary surface free energy to nucleate bubbles of the hadronic phase. The bubbles expand until the whole Universe is converted into the low temperature phase. An important parameter for the dynamics of the phase transition is the confined-deconfined interface tension. It sets the scale for the spatial size of a nucleating bubble, and hence it determines the size of spatial inhomogeneities produced during the phase transition. The inhomogeneities in turn may influence the primordial nucleosynthesis of light elements that takes place after the phase transition, because they induce fluctuations in the baryon density. In the pure gauge theory the dynamics of the phase transition is very different due to complete wetting. Then, above the transition temperature, a network of deconfined-deconfined domain walls is present. As one approaches the phase transition the domain walls split into pairs of confined-deconfined interfaces with a complete wetting layer of confined phase in between. In this scenario the creation of the confined phase does not require super-cooling, because the free energy of two confined-deconfined interfaces is the same as for one deconfined-deconfined domain wall. In fact, the confined phase appears already slightly above T_c . Furthermore, the spatial structure of the confined phase is determined by the domain wall network that emerged from the dynamics of the gluon plasma above T_c , and not by the value of the interface tension. In addition, also the usual bubble nucleation may occur, but it is certainly not the only mechanism by which the phases transform into each other when complete wetting occurs.

Since the confined-deconfined interface tension is an important parameter of the deconfinement phase transition, one would like to compute its value. Unlike the critical exponents of complete wetting the value of the interface tension is not universal. It does not follow from the effective Φ^4 model, and hence it must be calculated from the full $SU(3)$ gauge theory. This is possible in numerical lattice simulations of the theory.

7.4 Complete versus Incomplete Wetting

Above the deconfinement phase transition of the $SU(3)$ pure gauge theory a $\mathbb{Z}(3)$ symmetry is spontaneously broken, and three deconfined phases coexist with each other. Spatial regions of different phases are separated by domain walls with a

free energy

$$F = \sigma_{dd}(T)AT, \quad (7.4.1)$$

where $\sigma_{dd}(T)$ is the so-called reduced deconfined-deconfined interface tension, T is the temperature and A is the interface area. At very high temperatures the interface tension has been computed perturbatively

$$\sigma_{dd}(T_c) = \frac{8\pi^2 T^2}{9g}, \quad (7.4.2)$$

where g is the gauge coupling renormalized at the scale T . At the critical temperature T_c of the deconfinement phase transition in addition the confined phase arises, which is separated from the deconfined phases by confined-deconfined interfaces with a reduced interface tension σ_{cd} . Thermodynamical stability requires

$$\sigma_{dd}(T_c) \leq 2\sigma_{cd}. \quad (7.4.3)$$

For $\sigma_{dd}(T_c) > 2\sigma_{cd}$ a deconfined-deconfined domain wall could lower its free energy by splitting into two confined-deconfined interfaces, which would lead to

$$\sigma_{dd}(T_c) = 2\sigma_{cd}. \quad (7.4.4)$$

In this case one speaks of complete wetting, because a deconfined-deconfined interface is then always completely wet by the confined layer between the two confined-deconfined interfaces. In general, equilibrium of forces at a vertex where two deconfined phases are in contact with the confined phase implies

$$\sigma_{dd}(T_c) = 2\sigma_{cd} \cos \frac{\theta}{2}, \quad (7.4.5)$$

where θ is the opening angle of a lens of confined phase that forms at the deconfined-deconfined interface. When $\sigma_{dd}(T_c) < 2\sigma_{cd}$ the angle θ is different from zero. Then one speaks of incomplete or partial wetting, because the deconfined-deconfined interface is only partially wet by the confined phase. For complete wetting ($\sigma_{dd}(T_c) = 2\sigma_{cd}$) the angle θ vanishes and the lens degenerates to a confined layer that completely wets the deconfined-deconfined domain wall. Complete wetting is a critical phenomenon of interfaces that arises although the phase transition in the bulk is of first order and not critical. In particular, the thickness z_0 of the complete wetting layer grows to macroscopic size as one approaches the phase transition from above

$$z_0 \propto (T - T_c)^{-\psi}. \quad (7.4.6)$$

The divergence of z_0 is characterized by the critical exponent ψ . At the same time the order parameter (in our case the expectation value of the Polyakov loop) at the center of the interface

$$\Phi_1(0) \propto (T - T_c)^\beta \quad (7.4.7)$$

goes to zero with another critical exponent β . In the next section we will compute the critical exponents analytically and we will see that the values are

$$\psi = 0, \beta = \frac{1}{2}. \quad (7.4.8)$$

In particular, the thickness of the confined complete wetting layer z_0 diverges only logarithmically as one approaches the phase transition. The critical exponents will be computed using an effective 3-dimensional $\mathbb{Z}(3)$ symmetric Φ^4 theory for the Polyakov loop, which is in the same universality class (for critical interface effects) as the pure gauge theory.

7.5 An Effective 3-d $\mathbb{Z}(3)$ Symmetric Φ^4 Model

Let us again consider the Polyakov loop

$$\Phi(\vec{x}) = \text{Tr}(\mathcal{P} \exp \int_0^\beta dt G_4(\vec{x}, t)) \in \mathbb{C}. \quad (7.5.1)$$

It is invariant under ordinary periodic gauge transformations, but it transforms into

$$\Phi'(\vec{x}) = \Phi(\vec{x})z^*, \quad (7.5.2)$$

under transformations which are periodic only up to a center element $z \in \mathbb{Z}(3)$. Another important symmetry of the system is charge conjugation, which transforms the Polyakov loop into its complex conjugate

$$\Phi'(\vec{x}) = \Phi^*(\vec{x}). \quad (7.5.3)$$

The Polyakov loop is a 3-dimensional complex scalar field, whose dynamics is governed by the effective action

$$\exp(-S_{eff}[\Phi]) = \int \mathcal{D}G_\mu \delta[\Phi(\vec{x}) - \text{Tr}(\mathcal{P} \exp \int_0^\beta dt G_4(\vec{x}, t))] \exp(-S[G_\mu]). \quad (7.5.4)$$

The δ -functional ensures that Φ obeys eq.(7.2.2). The full effective action is certainly an unpleasant nonlocal functional, which is not well suited for further analytic investigations. When one is interested in universal critical phenomena — in this case in complete wetting — one may, however, take any action from the same universality class as $S_{eff}[\Phi]$. A pleasant local action is

$$S[\Phi] = \int d^3x [\frac{1}{2} \partial_i \Phi^* \partial_i \Phi + V(\Phi)]. \quad (7.5.5)$$

In order to belong to the same universality class as the full effective action the action S should at least have the same symmetries, i.e. it should be invariant under charge conjugation and against $\mathbb{Z}(3)$ transformations. The kinetic term is invariant under both. The potential term, however, must obey

$$V(\Phi^*) = V(\Phi), \quad V(\Phi z^*) = V(\Phi), \quad (7.5.6)$$

where $z \in \mathbb{Z}(3)$. For simplicity let us assume that $V(\Phi)$ is a polynomial in Φ of order at most Φ^4 . Decomposing the Polyakov loop into real and imaginary parts

$$\Phi = \Phi_1 + i\Phi_2, \quad (7.5.7)$$

and using the symmetry requirements of eq.(7.5.6) one obtains

$$V(\Phi) = a|\Phi|^2 + b\Phi_1(\Phi_1^2 - 3\Phi_2^2) + c|\Phi|^4. \quad (7.5.8)$$

Stability of the problem requires $c > 0$. For $a > 0$ the potential has a minimum at $\Phi_1 = \Phi_2 = 0$, which corresponds to the confined phase and which we denote by $\Phi^{(4)} = (0, 0)$. In addition, there may be minima corresponding to the three deconfined phases. They occur for $0 < a < 9b^2/32c$, $b < 0$ and they are located at

$$\begin{aligned} \Phi^{(1)} &= (\Phi_0, 0), \\ \Phi^{(2)} &= \left(-\frac{1}{2}\Phi_0, \frac{\sqrt{3}}{2}\Phi_0\right), \\ \Phi^{(3)} &= \left(-\frac{1}{2}\Phi_0, -\frac{\sqrt{3}}{2}\Phi_0\right), \end{aligned} \quad (7.5.9)$$

where

$$\Phi_0 = \frac{-3b + \sqrt{9b^2 - 32ac}}{8c}. \quad (7.5.10)$$

The phase transition occurs when all four minima are degenerate, i.e. when

$$V(\Phi^{(1)}) = V(\Phi^{(2)}) = V(\Phi^{(3)}) = V(\Phi^{(4)}). \quad (7.5.11)$$

In our parameterization this corresponds to $b^2 = 4ac$.

7.6 Interfaces and Critical Exponents

Let us consider the theory in a spatial volume with periodic boundary conditions in the x - and y -directions with lengths L_x and L_y , and with an infinite

z -direction. Then between different bulk phases interfaces arise, which are closed due to periodic boundary conditions. A planar interface perpendicular to the z -direction has the area $A = L_x L_y$ and it is described by an order parameter profile

$$\Phi(\vec{x}) = \Phi(z). \quad (7.6.1)$$

The form of $\Phi(z)$ follows from the classical equations of motion corresponding to the action $S[\Phi]$, which are given by

$$\begin{aligned} \Phi_1'' &= \frac{\partial V}{\partial \Phi_1} = \Phi_1(2a + 4c|\Phi|^2) + 3b(\Phi_1^2 - \Phi_2^2), \\ \Phi_2'' &= \frac{\partial V}{\partial \Phi_2} = \Phi_2(2a + 4c|\Phi|^2) - 6b\Phi_1\Phi_2. \end{aligned} \quad (7.6.2)$$

Here a prime denotes a derivative with respect to z . The interface action is related to the reduced interface tension by

$$S = L_x L_y \int dz(T + V) = 2A \int dzV = \sigma A. \quad (7.6.3)$$

Note that interfaces are topological excitations, like kinks or instantons, that interpolate between two different vacuum states (phases) of the model.

The simplest interface is the one between the confined and the deconfined phase, which exists only at the phase transition temperature corresponding to $b^2 = 4ac$. The interface interpolates between the confined phase $\Phi^{(4)}$ at large negative z and the deconfined phase $\Phi^{(1)}$ at large positive z . At the critical point the solution of the classical equations of motion gives

$$\Phi_1(z) = \frac{\Phi_0}{2}[1 + \tanh \frac{\mu}{2}(z - z_0)], \quad \Phi_2(z) = 0, \quad (7.6.4)$$

where $\mu = -b/\sqrt{2c}$ and $b < 0$. The corresponding action yields a reduced interface tension

$$\sigma_{cd} = S/A = \frac{1}{6}\mu\Phi_0^2. \quad (7.6.5)$$

Now we increase the temperature slightly above the phase transition, such that we are in the deconfined phase. Then also deconfined-deconfined interfaces arise. Such an interface separating the phases $\Phi^{(2)}$ and $\Phi^{(3)}$ can be constructed by combining two confined-deconfined interfaces, one centered around $z = z_0$ and the other one centered around $z = -z_0$

$$\begin{aligned} \Phi_1(z) &= -\frac{1}{4}\Phi_0[2 + \tanh \frac{\mu}{2}(z - z_0) - \tanh \frac{\mu}{2}(z + z_0)], \\ \Phi_2(z) &= \frac{\sqrt{3}}{4}\Phi_0[\tanh \frac{\mu}{2}(z - z_0) + \tanh \frac{\mu}{2}(z + z_0)]. \end{aligned} \quad (7.6.6)$$

Minimizing the action with respect to the free parameter z_0 — the thickness of the confined layer between the two confined-deconfined interfaces — one obtains

$$z_0 = -\frac{1}{2\mu} \ln\left(\frac{1}{2} - \frac{2ac}{b^2}\right). \quad (7.6.7)$$

The critical temperature corresponds to $b^2 = 4ac$. Then z_0 diverges logarithmically. This shows that at the critical point complete wetting occurs. Because of the logarithmic divergence the critical exponent is

$$\psi = 0. \quad (7.6.8)$$

For the expectation value of the Polyakov loop at the center of the interface one finds

$$\Phi_1(0) = -\Phi_0 \sqrt{\frac{1}{2} - \frac{2ac}{b^2}}, \quad (7.6.9)$$

which goes to zero as one approaches the phase transition. Because of the square root, the corresponding critical exponent is

$$\beta = \frac{1}{2}. \quad (7.6.10)$$

The same critical exponents occur for condensed matter systems with short range forces between different interfaces. For the gluon system one would also expect short range forces, because there are no massless excitations even in the plasma phase. The reduced deconfined-deconfined interface tension of the above solution is given by

$$\sigma_{dd}(T) = S/A = \mu \Phi_0^2 \left[\frac{1}{3} - 2\left(\frac{1}{2} - \frac{2ac}{b^2}\right) \right]. \quad (7.6.11)$$

Therefore

$$2\sigma_{cd} - \sigma_{dd}(T) = 2\mu \Phi_0^2 \left(\frac{1}{2} - \frac{2ac}{b^2}\right) \propto (T - T_c)^{1-\psi}, \quad (7.6.12)$$

in agreement with general scaling arguments. The analytic instanton solution from above assumes a dilute gas of confined-deconfined interfaces and treats interactions between interfaces only to lowest order. In general, the equations of motion can only be solved numerically. For example, one may consider the deconfined-deconfined interface between the phases $\Phi^{(2)}$ and $\Phi^{(3)}$ for any temperature above T_c .

We have seen that wetting is always complete in the effective Φ^4 model. Of course, this does not mean that complete wetting also occurs in the pure $SU(3)$ gauge theory. Such a question can only be decided in the gauge theory itself, for example by performing numerical simulations. Indeed, numerical simulations show that the pure $SU(3)$ gauge theory shows complete wetting.

7.7 The Electroweak Phase Transition

At temperatures in the 300 GeV range, i.e. at times around 10^{-12} sec after the Bang, a phase transition must have happened in the Electroweak sector of the standard model. At very early times the $SU(2)_L \otimes U(1)_Y$ gauge symmetry was not yet spontaneously broken to $U(1)_{em}$. Correspondingly, at that time all fermions and gauge bosons have been massless. Only the scalar field Φ had a mass. The order of the phase transition (and hence the strength of cosmological effects) depends on the mass of the Higgs particle via the temperature dependence of the effective potential of the scalar field. For a heavy Higgs particle (mass > 100 GeV) as it seems to exist in our Universe, the phase transition is second order, and cosmological effects are very weak, at least as long as the gauge couplings g , g' and g_s can be neglected. When one includes the effect of the gauge couplings, the phase transition becomes weakly first order.

In case of a light Higgs particle (mass ≈ 10 GeV), which is already ruled out experimentally, the phase transition would be strongly first order, and would lead to a strong super-cooling of the symmetric phase. This could have delayed the electroweak phase transition until the QCD transition. In that case the generated entropy would be unacceptably large, such that cosmological effects alone lead to a lower bound on the Higgs mass of ≈ 10 GeV.

Chapter 8

Grand Unified Theories

Although the standard model of particle physics agrees very well with experiments, from a theoretical point of view it is not completely satisfactory. In particular, the large number of free parameters like fermion masses (Yukawa couplings) or gauge boson masses (gauge couplings) suggests the existence of a more fundamental underlying theory, which would allow to explain these parameters. Also gravity is not included in the standard model, which again points to its incompleteness. Even electromagnetism and the weak force are not really unified in the standard model, because there are two independent coupling constants g and g' . QCD has yet another coupling constant g_s , which is unrelated to g and g' . In the framework of grand unified theories (GUT) one embeds the electroweak and strong interactions in one simple gauge group (e.g. $SU(5)$), which leads to a relation between g , g' and g_s .

Of course, the $SU(5)$ symmetry is too big, in order to be realized at low temperatures. It must be spontaneously broken to the $SU(3)_c \otimes SU(2)_L \otimes U(1)_Y$ symmetry of the standard model. In grand unified theories this happens at about 10^{14} GeV at about 10^{-34} sec after the Big Bang. At present (and in the foreseeable future) these energy scales cannot be probed experimentally. Hence, we must now rely on theoretical arguments, and sometimes on speculation.

The unification of the electroweak and strong interactions naturally leads to the decay of strongly interacting particles into particles which participate in the electroweak interactions only, for example, quarks can decay into leptons. This inevitably leads to the decay of the proton, and hence to a violation of baryon number conservation. Experimentally, the proton is extremely long lived (life-time $> 10^{32}$ years). This has already ruled out the simplest version of the

$SU(5)$ theory. Still, we will discuss that model in some detail, because it is the simplest representative in a larger class of GUTs. Other GUT theories allow larger life-times of the proton, and are not ruled out experimentally. On the other hand, proton decay is a very attractive feature of GUTs, because it may offer an explanation of the baryon asymmetry — the fact that our Universe consists of matter — not of anti-matter.

8.1 The minimal $SU(5)$ Model

The group $SU(n)$ has rank $n - 1$, i.e. $n - 1$ of the $n^2 - 1$ generators commute with each other. The rank of a simple $U(1)$ group is 1. Thus, the rank of the standard model group $SU(3)_c \otimes SU(2)_L \otimes U(1)_Y$ is $2 + 1 + 1 = 4$. Hence, if we want to embed that group in a simple Lie group, its rank must be at least 4. The simplest Lie group with that property is $SU(5)$, which has rank 4 and $5^2 - 1 = 24$ generators. Consequently, in an $SU(5)$ gauge theory there are 24 gauge bosons. When the standard model is embedded in $SU(5)$, half of the gauge bosons can be identified with known particles: $SU(3)$ has $3^2 - 1 = 8$ gluons, $SU(2)$ has $2^2 - 1 = 3$ W-bosons, and $U(1)$ has one B-boson, which, together with W^3 , forms the Z-boson and the photon. The remaining 12 gauge bosons of $SU(5)$ are new particles, which are called X and Y. To make these particles heavy, the $SU(5)$ symmetry must be spontaneously broken to $SU(3)_c \otimes SU(2)_L \otimes U(1)_Y$. Again, this is achieved via the Higgs mechanism, here with a scalar field transforming under the 24-dimensional adjoint representation of $SU(5)$. We write

$$\Phi(x) = \Phi^a(x)\eta_a, \quad a \in \{1, 2, \dots, 24\}. \quad (8.1.1)$$

The η_a are the 24 generators of $SU(5)$. They are Hermitean, traceless 5×5 matrices, analogous to the 3 Pauli matrices of $SU(2)$. Under gauge transformations $g \in SU(5)$ the scalar field transforms as

$$\Phi(x)' = g(x)\Phi(x)g(x)^\dagger. \quad (8.1.2)$$

We introduce a Φ^4 potential, which takes the form

$$V(\Phi) = \frac{1}{2}m^2\text{Tr}(\Phi^2) + \lambda_1(\text{Tr}(\Phi^2))^2 + \lambda_2\text{Tr}(\Phi^4). \quad (8.1.3)$$

The potential is gauge invariant due to the cyclic nature of the trace. We can choose a unitary gauge, in which the scalar field is diagonal (one uses the unitary

transformation $g(x)$ to diagonalize the Hermitean matrix $\Phi(x)$

$$\Phi = \begin{pmatrix} \Phi_1 & 0 & 0 & 0 & 0 \\ 0 & \Phi_2 & 0 & 0 & 0 \\ 0 & 0 & \Phi_3 & 0 & 0 \\ 0 & 0 & 0 & \Phi_4 & 0 \\ 0 & 0 & 0 & 0 & \Phi_5 \end{pmatrix}, \quad \Phi_i \in \mathbb{R}, \quad \sum_i \Phi_i = 0. \quad (8.1.4)$$

The potential then takes the form

$$V(\Phi) = \frac{1}{2} \sum_i \Phi_i^2 + \lambda_1 \left(\sum_i \Phi_i^2 \right)^2 + \lambda_2 \sum_i \Phi_i^4. \quad (8.1.5)$$

The minima of the potential are characterized by

$$\frac{\partial V}{\partial \Phi_i} = m^2 \Phi_i + 4\lambda_1 \sum_j \Phi_j^2 \Phi_i + 4\lambda_2 \Phi_i^3 = C. \quad (8.1.6)$$

Here C is a Lagrange multiplier that implements the constraint $\sum_i \Phi_i = 0$. We are interested in minima with an unbroken $SU(3)_c \otimes SU(2)_L \otimes U(1)_Y$ symmetry, for which

$$\Phi_1 = \Phi_2 = \Phi_3, \quad \Phi_4 = \Phi_5. \quad (8.1.7)$$

Hence, we can write

$$\Phi = v \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -3/2 & 0 \\ 0 & 0 & 0 & 0 & -3/2 \end{pmatrix}, \quad (8.1.8)$$

such that

$$\begin{aligned} m^2 v + 4\lambda_1 v^2 \left(3 + 2\frac{9}{4}\right) v + 4\lambda_2 v^3 &= C, \\ -\frac{3}{2} m^2 v - 4\lambda_1 v^2 \left(3 + 2\frac{9}{4}\right) \frac{3}{2} v - 4\lambda_2 v^3 \frac{27}{8} &= C \Rightarrow \\ C = \frac{4}{5} \lambda_2 v^3 \left(3 - \frac{27}{4}\right) &= -3\lambda_2 v^3 \Rightarrow \\ m^2 v + \lambda_1 30 v^3 + \lambda_2 7 v^3 &= 0 \Rightarrow v = \sqrt{-\frac{m^2}{30\lambda_1 + 7\lambda_2}}. \end{aligned} \quad (8.1.9)$$

The value of the potential at the minimum is given by

$$\begin{aligned}
V(\Phi) &= \frac{1}{2}m^2v^2(3 + 2\frac{9}{4}) + \lambda_1v^4(3 + 2\frac{9}{4})^2 + \lambda_2v^4(3 + 2\frac{81}{16}) \\
&= \frac{1}{2}m^2v^2\frac{15}{2} + \lambda_1v^4\frac{225}{4} + \lambda_2v^4\frac{105}{8} \\
&= -\frac{15}{4}v^4(30\lambda_1 + 7\lambda_2) + \lambda_1v^4\frac{225}{4} + \lambda_2v^4\frac{105}{8} \\
&= v^4(-\frac{225}{4}\lambda_1 - \frac{105}{8}\lambda_2) = -m^4\frac{15}{8}\frac{1}{30\lambda_1 + 7\lambda_2}. \tag{8.1.10}
\end{aligned}$$

For $\lambda_1, \lambda_2 > 0$ the value of the potential is negative, indicating that the $SU(5)$ symmetric phase at $\Phi = 0$ with $V(\Phi) = 0$ is not the true vacuum. It is instructive to convince oneself that other symmetry breaking patterns — for example to $SU(4) \otimes U(1)$ — are not dynamically preferred over $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$ breaking.

Let us now consider the gauge fields

$$V_\mu = ig_5 V_\mu^a(x) \eta_a. \tag{8.1.11}$$

Under non-Abelian gauge transformations we have

$$V_\mu(x)' = g(x)(V_\mu(x) + \partial_\mu)g(x)^\dagger. \tag{8.1.12}$$

For an adjoint Higgs field the covariant derivative takes the form

$$D_\mu \Phi(x) = \partial_\mu \Phi(x) + [V_\mu(x), \Phi(x)]. \tag{8.1.13}$$

It is instructive to show that this indeed transforms covariantly. Introducing the field strength tensor

$$V_{\mu\nu} = \partial_\mu V_\nu(x) - \partial_\nu V_\mu(x) + [V_\mu(x), V_\nu(x)], \tag{8.1.14}$$

the bosonic part of the $SU(5)$ GUT Lagrange density takes the form

$$\mathcal{L}(\Phi, V_\mu) = \frac{1}{2}\text{Tr} D^\mu \Phi D_\mu \Phi - V(\Phi) - \frac{1}{4}\text{Tr} V^{\mu\nu} V_{\mu\nu}. \tag{8.1.15}$$

Next we insert the vacuum value of the scalar field to obtain the mass terms for the gauge field

$$\frac{1}{2}\text{Tr} D^\mu \Phi D_\mu \Phi = \text{Tr}[V^\mu, \Phi][V_\mu, \Phi]. \tag{8.1.16}$$

We introduce the X and Y-bosons via

$$V_\mu = \begin{pmatrix} & G_\mu & & X_\mu^r & Y_\mu^r \\ & & & X_\mu^g & Y_\mu^g \\ & & & X_\mu^b & Y_\mu^b \\ X_\mu^{r*} & X_\mu^{g*} & X_\mu^{b*} & & \\ Y_\mu^{r*} & Y_\mu^{g*} & Y_\mu^{b*} & & W_\mu \end{pmatrix}. \quad (8.1.17)$$

X and Y-bosons are color triplets and electroweak doublets. They are the fields that become massive after the spontaneous breakdown of $SU(5)$ to $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$, because one obtains

$$\begin{aligned} [V_\mu, \Phi] &= v \begin{pmatrix} & G_\mu \mathbb{1} & & -\frac{3}{2}X_\mu^r & -\frac{3}{2}Y_\mu^r \\ & & & -\frac{3}{2}X_\mu^g & -\frac{3}{2}Y_\mu^g \\ & & & -\frac{3}{2}X_\mu^b & -\frac{3}{2}Y_\mu^b \\ X_\mu^{r*} & X_\mu^{g*} & X_\mu^{b*} & & \\ Y_\mu^{r*} & Y_\mu^{g*} & Y_\mu^{b*} & & -\frac{3}{2}W_\mu \mathbb{1} \end{pmatrix} \\ &= v \begin{pmatrix} & G_\mu \mathbb{1} & & X_\mu^r & Y_\mu^r \\ & & & X_\mu^g & Y_\mu^g \\ & & & X_\mu^b & Y_\mu^b \\ -\frac{3}{2}X_\mu^{r*} & -\frac{3}{2}X_\mu^{g*} & -\frac{3}{2}X_\mu^{b*} & & \\ -\frac{3}{2}Y_\mu^{r*} & -\frac{3}{2}Y_\mu^{g*} & -\frac{3}{2}Y_\mu^{b*} & & -\frac{3}{2}W_\mu \mathbb{1} \end{pmatrix} \\ &= v \begin{pmatrix} & G_\mu \mathbb{1} & & -\frac{5}{2}X_\mu^r & -\frac{5}{2}Y_\mu^r \\ & & & -\frac{5}{2}X_\mu^g & -\frac{5}{2}Y_\mu^g \\ & & & -\frac{5}{2}X_\mu^b & -\frac{5}{2}Y_\mu^b \\ -\frac{5}{2}X_\mu^{r*} & -\frac{5}{2}X_\mu^{g*} & -\frac{5}{2}X_\mu^{b*} & & \\ -\frac{5}{2}Y_\mu^{r*} & -\frac{5}{2}Y_\mu^{g*} & -\frac{5}{2}Y_\mu^{b*} & & 0 \end{pmatrix}, \end{aligned} \quad (8.1.18)$$

and hence

$$\text{Tr}[V^\mu, \Phi][V_\mu, \Phi] = -\frac{25}{2}v^2(X^{\mu*}X_\mu + Y^{\mu*}Y_\mu). \quad (8.1.19)$$

The X and Y-fields thus pick up the mass

$$m_X^2 = m_Y^2 = \frac{25}{2}v^2 g_5^2. \quad (8.1.20)$$

The 12 gauge bosons become massive by eating 12 Goldstone bosons. Indeed, due to the Goldstone theorem, in this case there are $24 - 8 - 3 - 1 = 12$ Goldstone bosons.

8.2 The Fermion Multiplets

How can we arrange quarks and leptons in representations of $SU(5)$? Let us consider the fermions of the first generation

$$\begin{pmatrix} \nu_{eL} \\ e_L \end{pmatrix}, e_R, \begin{pmatrix} u_L^r \\ d_L^r \end{pmatrix}, u_R^r, d_R^r, \begin{pmatrix} u_L^g \\ d_L^g \end{pmatrix}, u_R^g, d_R^g, \begin{pmatrix} u_L^b \\ d_L^b \end{pmatrix}, u_R^b, d_R^b. \quad (8.2.1)$$

These are 15 fermionic degrees of freedom. The fundamental representation of $SU(5)$ is 5-dimensional. It decomposes into an $SU(3)$ triplet and an $SU(2)$ doublet with corresponding $U(1)$ quantum numbers

$$\{5\} = \{3, 1\}_{-2/3} \oplus \{1, 2\}_1. \quad (8.2.2)$$

Let us couple two fundamental representations in $SU(2)$, $SU(3)$ and $SU(5)$

$$\begin{aligned} \{2\} \otimes \{2\} &= \{3\} \oplus \{1\}, \\ \{3\} \otimes \{3\} &= \{6\} \oplus \{\bar{3}\}, \\ \{5\} \otimes \{5\} &= \{15\} \oplus \{10\}. \end{aligned} \quad (8.2.3)$$

Indeed, there is a 15-dimensional representation. Can we host the fermions of a generation in that representation? We investigate the $SU(3)$ and $SU(2)$ content

$$\begin{aligned} \{5\} \otimes \{5\} &= (\{3, 1\}_{-2/3} \oplus \{1, 2\}_1) \otimes (\{3, 1\}_{-2/3} \oplus \{1, 2\}_1) \\ &= \{3, 1\}_{-2/3} \otimes \{3, 1\}_{-2/3} \oplus \{1, 2\}_1 \otimes \{3, 1\}_{-2/3} \\ &\oplus \{3, 1\}_{-2/3} \otimes \{1, 2\}_1 \oplus \{1, 2\}_1 \otimes \{1, 2\}_1 \\ &= \{6, 1\}_{-4/3} \oplus \{\bar{3}, 1\}_{-4/3} \oplus \{3, 2\}_{1/3} \oplus \{3, 2\}_{1/3} \\ &\oplus \{1, 3\}_2 \oplus \{1, 1\}_2 = \{15\} \oplus \{10\}. \end{aligned} \quad (8.2.4)$$

The symmetric combination $\{15\}$ contains

$$\{15\} = \{6, 1\}_{-4/3} \oplus \{3, 2\}_{1/3} \oplus \{1, 3\}_2, \quad (8.2.5)$$

while the anti-symmetric combination $\{10\}$ is

$$\{10\} = \{\bar{3}, 1\}_{-4/3} \oplus \{3, 2\}_{1/3} \oplus \{1, 1\}_2. \quad (8.2.6)$$

The standard model does not contain fermions in a sextet representation of $SU(3)$ (the quarks are triplets and the leptons are singlets). Hence the 15 fermions of a generation do not form a 15-plet of $SU(5)$. We have

$$\begin{aligned} \left\{ \begin{pmatrix} u_L^r \\ d_L^r \end{pmatrix}, \begin{pmatrix} u_L^g \\ d_L^g \end{pmatrix}, \begin{pmatrix} u_L^b \\ d_L^b \end{pmatrix} \right\} &= \{3, 2\}_{1/3}, \quad \{u_R^r, u_R^g, u_R^b\} = \{3, 1\}_{4/3}, \\ \{d_R^r, d_R^g, d_R^b\} &= \{3, 1\}_{-2/3}, \quad \begin{pmatrix} \nu_{eL} \\ e_L \end{pmatrix} = \{1, 2\}_{-1}, \quad e_R = \{1, 1\}_{-2}. \end{aligned} \quad (8.2.7)$$

Both $\{3, 2\}$ and $\{1, 1\}$ are contained in $\{10\}$. However, there is also a $\{\bar{3}, 1\}$. Furthermore, we cannot mix left and right-handed fermion components. We use charge conjugation to write everything in terms of left-handed fields

$$\begin{aligned} & \left\{ \begin{pmatrix} u_L^r \\ d_L^r \end{pmatrix}, \begin{pmatrix} u_L^g \\ d_L^g \end{pmatrix}, \begin{pmatrix} u_L^b \\ d_L^b \end{pmatrix} \right\} = \{3, 2\}_{1/3}, \\ & \{^C u_R^r, ^C u_R^g, ^C u_R^b\} = \{\bar{3}, 1\}_{-4/3}, \quad \{^C d_R^r, ^C d_R^g, ^C d_R^b\} = \{\bar{3}, 1\}_{2/3}, \\ & \begin{pmatrix} \nu_{eL} \\ e_L \end{pmatrix} = \{1, 2\}_{-1}, \quad e_R = \{1, 1\}_2. \end{aligned} \quad (8.2.8)$$

Now we can identify a decouplet

$$\begin{aligned} & \left\{ \begin{pmatrix} u_L^r \\ d_L^r \end{pmatrix}, \begin{pmatrix} u_L^g \\ d_L^g \end{pmatrix}, \begin{pmatrix} u_L^b \\ d_L^b \end{pmatrix} \right\} \oplus \{^C u_R^r, ^C u_R^g, ^C u_R^b\} \oplus ^C e_R = \\ & \{3, 2\}_{1/3} \oplus \{\bar{3}, 1\}_{-4/3} \oplus \{1, 1\}_2 = \{10\}. \end{aligned} \quad (8.2.9)$$

The remaining five fermions are

$$\{^C d_R^r, ^C d_R^g, ^C d_R^b\} = \{\bar{3}, 1\}_{2/3}, \quad \begin{pmatrix} \nu_{eL} \\ e_L \end{pmatrix} = \{1, 2\}_{-1}. \quad (8.2.10)$$

They naturally fit into an anti-quintet

$$\{\bar{5}\} = \{\bar{3}, 1\}_{2/3} \oplus \{1, \bar{2}\}_{-1}. \quad (8.2.11)$$

In $SU(2)$ the representations $\{2\}$ and $\{\bar{2}\}$ are equivalent, such that

$$\{^C d_R^r, ^C d_R^g, ^C d_R^b\} \oplus \begin{pmatrix} \nu_{eL} \\ e_L \end{pmatrix} = \{\bar{3}, 1\}_{2/3} \oplus \{1, 2\}_{-1} = \{\bar{5}\}. \quad (8.2.12)$$

Hence, the fermions of one generation form a reducible 15-dimensional representation of $SU(5)$, which decomposes into $\{10\}$ and $\{\bar{5}\}$. In $SU(5)$ quarks and leptons are in the same irreducible representation. This immediately explains why proton and positron have the same electric charge — a fact that remains unexplained in the standard model.

Let us also determine the $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$ properties of the X and Y-bosons. They transform under the adjoint representation — the 24-plet, that is obtained from

$$\{5\} \otimes \{\bar{5}\} = \{24\} \oplus \{1\}. \quad (8.2.13)$$

On the other hand, we have

$$\begin{aligned}
\{5\} \otimes \{\bar{5}\} &= (\{3, 1\}_{-2/3} \oplus \{1, 2\}_1) \otimes (\{\bar{3}, 1\}_{2/3} \oplus \{1, 2\}_{-1}) \\
&= \{3, 1\}_{-2/3} \otimes \{\bar{3}, 1\}_{2/3} \oplus \{1, 2\}_1 \otimes \{\bar{3}, 1\}_{2/3} \\
&\oplus \{3, 1\}_{-2/3} \otimes \{1, 2\}_{-1} \oplus \{1, 2\}_1 \otimes \{1, 2\}_{-1} \\
&= \{8, 1\}_0 \oplus \{1, 1\}_0 \oplus \{\bar{3}, 2\}_{5/3} \oplus \{3, 2\}_{-5/3} \oplus \{1, 3\}_0 \oplus \{1, 1\}_0.
\end{aligned} \tag{8.2.14}$$

Hence, we identify

$$\begin{aligned}
\{24\} &= \{8, 1\}_0 \oplus \{1, 3\}_0 \oplus \{1, 1\}_0 \oplus \{\bar{3}, 2\}_{5/3} \oplus \{3, 2\}_{-5/3} \\
&= \{\text{Gluons}\} \oplus \{\text{W-bosons}\} \oplus \{\text{B-boson}\} \oplus \{\text{X, Y-bosons}\}.
\end{aligned} \tag{8.2.15}$$

The X and Y-bosons are $SU(3)$ triplets, $SU(2)$ doublets, and have hypercharge $5/3$.

Up to this point all fermions and all gauge bosons of the standard model are still massless. Of course, we could write Yukawa couplings and try to give mass to the fermions by coupling them to the adjoint scalar field. However, it turns out that this is inconsistent with the group theory of $SU(5)$. In principle, this is good news, because otherwise the fermions would naturally get masses at the GUT scale. However, this also means that we will need more scalar fields, in particular the $SU(2)$ complex doublet of the standard model, which is naturally contained in the fundamental representation of $SU(5)$. Hence, we introduce a 5-plet of scalar fields

$$H = \begin{pmatrix} H_1 \\ H_2 \\ H_3 \\ H_4 \\ H_5 \end{pmatrix}, \quad H_4 = \Phi_+, \quad H_5 = \Phi_0. \tag{8.2.16}$$

Now we can write down various Yukawa couplings. The group theory of $SU(5)$ implies

$$\{\bar{5}\} \otimes \{10\} = \{5\} \oplus \{45\}, \quad \{10\} \otimes \{10\} = \{\bar{5}\} \oplus \{\bar{45}\} \oplus \{50\}. \tag{8.2.17}$$

Hence, the fermions in the $\{\bar{5}\}$ and $\{10\}$ representations can be coupled to $\{5\}$ and $\{\bar{5}\}$ — i.e. together with the scalar 5-plet they can be coupled in a gauge invariant way (form an $SU(5)$ singlet).

8.3 Predictions of Grand Unified Theories

In the $SU(5)$ GUT there is only one gauge coupling g_5 . Hence, it must be possible to relate the standard model gauge couplings g , g' and g_s to that coupling. In an $SU(5)$ symmetric phase one has

$$g = g_s = g_5, \quad g' = \sqrt{\frac{3}{5}}g_5. \quad (8.3.1)$$

Hence, the Weinberg angle takes the form

$$\sin^2 \theta_W = \frac{g'^2}{g^2 + g'^2} = \frac{3g_5^2}{5g_5^2 + 3g_5^2} = \frac{3}{8}. \quad (8.3.2)$$

Furthermore, the structure of the Yukawa couplings implies

$$m_d = m_e, \quad m_s = m_\mu, \quad m_b = m_\tau. \quad (8.3.3)$$

This is not in agreement with experiments. However, we do not live in an $SU(5)$ symmetric world. One can use the renormalization group to run the above relations from the GUT scale, where they apply, down to our low energy scales. One obtains realistic values for particle masses and coupling constants when one puts the GUT scale at about $v = 10^{15}$ GeV. The masses of the X and Y-bosons are also in that range. The GUT scale is significantly below the Planck scale 10^{19} GeV, which justifies neglecting gravity in the above considerations.

An important prediction of GUTs is the instability of the proton. Proton decay proceeds via the X and Y-boson channel and is hence suppressed with the large mass of these particles. The following decays are possible

$$p \rightarrow e^+ + \pi^0, \quad p \rightarrow \bar{\nu}_e + \pi^+. \quad (8.3.4)$$

The resulting lifetime of the proton is given by

$$\tau_p \propto \frac{m_X^4}{m_p^5} \approx 10^{32} \text{ years}. \quad (8.3.5)$$

Although the predicted lifetime is much larger than the age of the Universe, the minimal $SU(5)$ model has been ruled out experimentally, because the proton indeed lives longer than the model predicts. More complicated GUTs based on $SO(10)$ or E_6 are not yet ruled out experimentally.

Chapter 9

Baryon Asymmetry

Why is there a baryon asymmetry in the Universe? In other words, why does the Universe consist of matter, and not also of anti-matter? The ratio of baryon to photon density $n_B/n_\gamma \approx 10^{-10}$ is an initial condition of the standard Big Bang model, which cannot be derived within it. Grand unified theories, however, allow us to compute a baryon asymmetry based on the presence of baryon number violating processes (like proton decay). This alone does not guarantee a baryon asymmetry. Besides baryon number, also C and CP must be violated, and the Universe must get out of thermal equilibrium. All these additional conditions are indeed satisfied in our Universe. Since the standard model is a chiral gauge theory, C is maximally violated, and also CP is explicitly broken (at least weakly). Furthermore, the expansion of the Universe implies that a true thermal equilibrium is never achieved. The Universe cools down and certain processes (for example those that violate baryon number) get out of equilibrium.

9.1 Evidence for a Baryon Asymmetry

Matter and anti-matter annihilate each other, for example, into photons. Regions in the Universe, in which matter and anti-matter systems collide, should thus be a source of very intense x-ray radiation. Such a thing has not been observed, indicating that the Universe consists of matter only, and not also of anti-matter. This is confirmed by the composition of cosmic rays, which contain 10000 times more protons than anti-protons, and even those few are due to secondary processes.

Of course, in the early Universe anti-matter has also been present. In particu-

lar, at temperatures above the GeV range, about the same number of quarks and anti-quarks have been around. During cooling quarks and anti-quarks annihilated each other, and all matter in the Universe today is a tiny fraction that survived the mass extinction. Comparing observed abundances of light nuclei with theoretical calculations of primordial nucleosynthesis leads to $n_B/n_\gamma \approx 10^{-10}$.

Since our Universe is electrically neutral, it contains as many electrons as protons, but almost no positrons. Correspondingly, there should also be a lepton asymmetry. However, since also the weakly interacting neutrinos contribute to it, it can at present not be detected experimentally.

9.2 Necessary Conditions for a Baryon Asymmetry

A trivial condition for the explanation of the baryon asymmetry is the existence of baryon number violating processes. Only then an initial state with unknown baryon number (for example $B = 0$), may turn into the situation that we observe now. GUTs indeed give rise to these processes. At temperatures in the 10^{14} GeV range they occur frequently, while today they are extremely suppressed. In the ultra-early Universe baryon number violating processes have been in thermal equilibrium, such that the effects of even earlier initial conditions are eliminated. In the $SU(5)$ theory the following processes are possible

$$\begin{aligned} u + u &\rightarrow X \rightarrow e^+ + \bar{d}, \\ u + d &\rightarrow Y \rightarrow \bar{\nu}_e + \bar{d}, \\ u + d &\rightarrow Y \rightarrow \bar{u} + e^+. \end{aligned} \tag{9.2.1}$$

All these processes violate baryon number ($\Delta B = 1$) and lepton number ($\Delta L = 1$) but $B - L$ remains unchanged. Indeed, $B - L$ is conserved in the $SU(5)$ GUT (but not necessarily in other GUTs), and it is also conserved in the standard model.

If the theory were C or CP invariant, baryon number violating processes would generate anti-baryons at the same rate as baryons, and no net baryon asymmetry could be generated. The charge conjugate of a left-handed quark is a right-handed anti-quark, i.e. baryon number changes sign under charge conjugation. Similarly, under CP a left-handed quark is turned into a left-handed anti-quark, and again baryon number changes sign. Both the standard model and the $SU(5)$ GUT are chiral gauge theories, in which C is maximally violated. In the standard model CP is weakly broken ($K - \bar{K}$ system), which manifests itself in the complex

phase of the Kobayashi-Maskawa matrix. This is possible only with at least three generations. The situation in the $SU(5)$ GUT is similar.

In thermal equilibrium baryon number violating processes proceed in both directions, i.e. baryons are generated but also annihilated. Hence, a net baryon asymmetry cannot be generated in thermal equilibrium. Formally, this results from CPT invariance (a symmetry of all consistent quantum field theories). Since the Universe expands and cools, baryon number violating processes get out of equilibrium, such that indeed all necessary conditions for the generation of a baryon asymmetry are satisfied in our Universe.

At very high temperatures ($\gg 10^{14}$ GeV) baryon number violating processes are in thermal equilibrium and $n_X = n_Y = n_\gamma$. When the temperature drops below m_X the X and Y -bosons get out of thermal equilibrium — they decay into quarks and leptons, but are not re-generated. Let us denote the rate for $X \rightarrow q\bar{l}$ by r . Then the rate for $X \rightarrow qq$ is $1 - r$. Correspondingly, the rate for $\bar{X} \rightarrow ql$ is \bar{r} , and the rate for $\bar{X} \rightarrow \bar{q}\bar{q}$ is $1 - \bar{r}$. Under the decay of an X and an \bar{X} one generates the baryon asymmetry

$$\Delta B = -\frac{1}{3}r + \frac{2}{3}(1 - r) + \frac{1}{3}\bar{r} - \frac{2}{3}(1 - \bar{r}) = \bar{r} - r = \varepsilon, \quad (9.2.2)$$

where ε is a measure of CP violation. The entropy density of the Universe is

$$s = \frac{2\pi^2}{45}g_*T^3, \quad (9.2.3)$$

where g_* is the number of relativistic degrees of freedom — for a typical GUT a number between 100 and 1000. The number density of X and \bar{X} bosons is

$$n_X = g_X \frac{1}{\pi^2} \zeta(3) T^3, \quad (9.2.4)$$

where g_X counts the degrees of freedom of X and \bar{X} bosons. After the decay of these particles a baryon asymmetry

$$n_B = n_X \varepsilon = g_X \frac{1}{\pi^2} \zeta(3) \frac{45}{2\pi^2} \frac{s}{g_*} \varepsilon \quad (9.2.5)$$

is generated. During the expansion of the Universe both the entropy and the baryon number are conserved. Today the entropy is almost entirely in the cosmic background radiation (as well as in the neutrinos, in case they are massless) and we can estimate

$$s = \frac{2\pi^2}{45} \pi^2 \frac{1}{\zeta(3)} n_\gamma, \quad (9.2.6)$$

such that today the baryon-photon ratio is

$$\frac{n_B}{n_\gamma} = \frac{g_X}{g_*} \varepsilon. \quad (9.2.7)$$

The observed ratio is $n_B/n_\gamma = 10^{-10}$, such that we need $\varepsilon \approx 10^{-8}$. This is not an unnatural number for typical GUTs. On the other hand, one has some free parameters that allow to adjust the right amount of baryon asymmetry. The predictive power of GUTs for the baryon asymmetry is therefore limited. Also the above arguments are only qualitatively correct. A more detailed investigation would require the numerical solution of rate equations similar to nucleosynthesis calculations.

9.3 Baryon Number Violation in the Standard Model

The classical Lagrange density of the standard model does not contain baryon number violating interactions. However, this does not imply that the standard model conserves baryon number after quantization. Indeed, due to the chiral couplings of the fermions, the baryon number current has an anomaly in the standard model. Although the Lagrange density has a global $U(1)$ baryon number symmetry, this symmetry is explicitly broken in the quantum theory. The same is true for lepton number. The difference, $B - L$, on the other hand, remains conserved. The existence of baryon number violating processes at the electroweak scale may change the baryon asymmetry that has been generated at the GUT scale.

Let us consider the vacuum structure of a non-Abelian gauge theory (like the $SU(2)$ sector of the standard model). A classical vacuum solution is

$$\Phi(\vec{x}) = \begin{pmatrix} \Phi_+(\vec{x}) \\ \Phi_0(\vec{x}) \end{pmatrix} = \begin{pmatrix} 0 \\ v \end{pmatrix}, \quad A_i(\vec{x}) = 0. \quad (9.3.1)$$

Of course, gauge transformations of this solution are also vacua. However, states that are related by a gauge transformation are physically equivalent, and one should not consider the other solutions as additional vacua. Still, there is a subtlety, because there are gauge transformations with different topological properties. First of all, there are the so-called small gauge transformations, which can be continuously deformed into the identity, and one should indeed not distinguish between states related by small gauge transformations. However, there are also large gauge transformations — those that can not be deformed into a trivial

gauge transformation — and they indeed give rise to additional vacuum states. The gauge transformations

$$g : \mathbb{R}^3 \rightarrow SU(2) \quad (9.3.2)$$

can be viewed as mappings from coordinate space into the group space. When one identifies points at spatial infinity \mathbb{R}^3 is compactified to S^3 . On the other hand, the group space of $SU(2)$ is also S^3 . Hence, the gauge transformations are mappings

$$g : S^3 \rightarrow S^3. \quad (9.3.3)$$

Such mappings are known to fall into topologically distinct classes characterized by a winding number

$$n[g] \in \Pi_3[SU(2)] = \mathbb{Z} \quad (9.3.4)$$

from the third homotopy group of the gauge group. In this case, mappings with any integer winding number are possible. Denoting a mapping with winding number n by g_n we can thus construct a set of topologically inequivalent vacuum states

$$\Phi^{(n)}(\vec{x}) = g_n(\vec{x}) \begin{pmatrix} 0 \\ v \end{pmatrix}, \quad A_i^{(n)}(\vec{x}) = g_n(\vec{x}) \partial_i g_n(\vec{x})^\dagger. \quad (9.3.5)$$

Topologically distinct vacua are separated by energy barriers, and thus there is a periodic potential in the space of field configurations.

Classically, the system is in one of the degenerate vacuum states. Quantum mechanically, however, the system can tunnel from one vacuum to another. It turns out that a transition from the vacuum (m) to the vacuum (n) is accompanied by a baryon number violating process of strength $\Delta B = N_g(n - m)$, where N_g is the number of generations of quarks and leptons. Also the lepton number changes by $\Delta L = N_g(n - m)$, such that $B - L$ is conserved. The tunnel amplitude — and hence the rate of baryon number violating processes — is controlled by the barrier height between adjacent classical vacua. The unstable field configuration at the top of the barrier is known as a sphaleron (meaning ready to decay). In the standard model the height of the barrier (the sphaleron energy) is given by $4\pi v/g$ and the resulting tunneling rate is

$$\exp\left(-\frac{8\pi^2}{g^2}\right) \approx \exp(-200), \quad (9.3.6)$$

which is totally negligible. Hence, for some time people assumed that baryon number violation in the standard model is only of academic interest. However, it was overlooked that in the early Universe one need not tunnel through the barrier — one can simply step over it classically due to large thermal fluctuations. Then one must assume that in the TeV range baryon number violating processes are

un-suppressed in the standard model. This means that any pre-existing baryon asymmetry — carefully created at the GUT scale — will be washed out, because baryon number violating processes are again in thermal equilibrium. Since the electroweak phase transition is of second or of weakly first order, it is unlikely (but not excluded) that a sufficient baryon asymmetry is re-generated at the electroweak scale.

However, we should not forget that $B - L$ is conserved in the standard model. This means that this mode is not thermalized. When baryon and lepton asymmetries ΔB_i and ΔL_i have been initially generated at the GUT scale, equilibrium sphaleron processes will imply that finally

$$\Delta(B_f + L_f) = 0, \quad (9.3.7)$$

but still

$$\Delta(B_f - L_f) = \Delta(B_i - L_i) = 0. \quad (9.3.8)$$

Hence, the present baryon and lepton asymmetries then are

$$\Delta B_f = -\Delta L_f = \frac{1}{2}\Delta(B_i - L_i). \quad (9.3.9)$$

This again leads to a problem, because also the minimal $SU(5)$ model conserves $B - L$. An asymmetry $\Delta(B_i - L_i)$ must hence be due to processes in the even earlier Universe. Then we would know as much as before. Fortunately, there is a way out. Other GUTs like $SO(10)$ and E_6 are not ruled out via proton decay and indeed do not conserve $B - L$. The reason for $B - L$ violation in these models is related to the existence of massive neutrinos. The so-called “see-saw” mechanism gives rise to one heavy neutrino of mass 10^{14} GeV and one light neutrino of mass in the eV range, that is identified with the neutrinos that we observe. Hence, we can explain the baryon asymmetry using GUTs only if the neutrinos are massive. Otherwise, we must assume that it was generated at times before 10^{-34} sec after the Big Bang, or we must find a way to go sufficiently out of thermal equilibrium around the electroweak phase transition and generate the baryon asymmetry via sphaleron processes.

Chapter 10

Topological Excitations

As we have discussed for the broken $\mathbb{Z}(3)$ symmetry in the pure gluon system, the spontaneous breakdown of a discrete symmetry leads to the presence of domain walls. Domain walls are 2-dimensional topological excitations, which are topologically stable, because the vacuum manifold is not simply connected.

Also when a continuous symmetry breaks spontaneously topological excitations may arise. In a type II superconductor, for example, the $U(1)$ gauge symmetry of electromagnetism is spontaneously broken. When the superconductor is placed in an external magnetic field, magnetic flux tubes — so-called Abrikosov strings — are formed, which are 1-dimensional topological excitations. Whenever a $U(1)$ symmetry (global or local) is spontaneously broken, this phenomenon arises. In the framework of particle physics this was first discussed by Nielsen and Olesen. The corresponding 1-dimensional topological excitations are therefore known as Nielsen-Olesen strings. When a $U(1)$ symmetry is spontaneously broken in the early Universe, 1-dimensional topological excitations arise, which may pass through the whole Universe. Networks of these so-called cosmic strings are discussed as a seed for galaxy formation. Cosmic strings also act as gravitational lenses, and they have been used in Gedanken experiments related to time machines. Still, cosmic strings are purely theoretical objects, which have not yet been observed in the sky.

Finally, also 0-dimensional (i.e. point-like) topological excitations may arise, which turn out to be magnetic monopoles. Such objects were first discussed by Dirac. He introduced magnetic monopoles by hand into electromagnetism. One has the freedom to choose $\vec{\nabla} \cdot \vec{B} = 0$, as Maxwell did, but one can also have $\vec{\nabla} \cdot \vec{B} \neq 0$. In GUTs, on the other hand, the presence of magnetic monopoles is

unavoidable. The mass of these so-called 't Hooft-Polyakov monopoles is naturally at the GUT scale. At the GUT phase transition a large number of monopoles is created in the early Universe. Since monopoles are topologically stable, they should still be around in the Universe today. However, there is no experimental evidence for the presence of monopoles. At the end of his life even Dirac did not believe in the existence of monopoles any more. GUTs thus pose the problem what happened to the monopoles in the Universe. The idea of inflation resolves this puzzle. When the Universe expands exponentially for some time after the GUT phase transition, the monopoles are diluted so much that they should be extremely rare in the Universe today. For topological reasons the electroweak phase transition is not associated with topologically stable excitations. Such objects — if they exist — would hence be relics of the very early Universe.

10.1 Domain Walls

Since we have discussed domain walls in some detail when we discussed the gluon system, we will be very brief here. We will just discuss some topological arguments, which guarantee the topological stability of such configurations. When a discrete symmetry is spontaneously broken in the early Universe, it is natural to expect the generation of domain walls. This is because then there is a set of degenerate vacua (three in the case of $\mathbb{Z}(3)$ breaking in the gluon system)

$$\mathcal{M} = \{\Phi | \Phi \text{ is a minimum of } V(\Phi)\}. \quad (10.1.1)$$

Let us now consider two halves of the Universe (left $z < 0$, right $z > 0$). At $z = -\infty$ and at $z = \infty$ the field Φ must assume a vacuum value in order to have finite energy, i.e.

$$\Phi(z = -\infty), \Phi(z = \infty) \in \mathcal{M}. \quad (10.1.2)$$

Topologically, the points $z = -\infty$ and $z = \infty$ correspond to a 0-dimensional sphere $\{-\infty, \infty\} = S^0$, such that the field at spatial infinity can be viewed as a map

$$\Phi : S^0 \rightarrow \mathcal{M}. \quad (10.1.3)$$

Again, such mappings are characterized by homotopy classes. The relevant homotopy group is $\Pi_0(\mathcal{M})$. It is non-trivial only if the vacuum manifold is not simply connected. For the system of hot gluons, for example, $\mathcal{M} = \mathbb{Z}(3)$ and

$$\Pi_0(\mathcal{M}) = \mathbb{Z}(3). \quad (10.1.4)$$

In the standard model, on the other hand, one has a continuous symmetry and \mathcal{M} is simply connected. Hence

$$\Pi_0(\mathcal{M}) = \{0\}, \quad (10.1.5)$$

and no domain walls arise. This is fine, because domain walls pose a problem for cosmology. Since their energy is proportional to their area, and since they would span the whole Universe, domain walls carry enormous energies. They would easily close the Universe, and would be in contradiction with cosmological observations.

10.2 Cosmic Strings

Let us consider scalar electrodynamics as the simplest model with cosmic strings. We can imagine that this theory occurs as some extension of the standard model (for example, as part of a GUT theory). Before we introduce gauge fields, the model has a global $U(1)$ symmetry, which is related to phase transformations of the complex scalar field $\Phi(x) \in \mathbb{C}$. The corresponding Lagrange density is

$$\mathcal{L}(\Phi, \partial_\mu \Phi) = \partial^\mu \Phi^* \partial_\mu \Phi - V(\Phi), \quad V(\Phi) = m^2 |\Phi|^2 + \lambda |\Phi|^4. \quad (10.2.1)$$

When $m^2 < 0$ the $U(1)$ symmetry gets spontaneously broken, and we obtain one massless Goldstone boson. Hence, we should not identify this $U(1)$ symmetry with the one of electromagnetism, because that symmetry is not spontaneously broken (at least not outside superconducting matter).

Let us now consider cylindrically symmetric field configurations

$$\Phi(x) = \Phi(\vec{x}, t) = \Phi(\vec{x}) = \Phi(\rho, \varphi), \quad (10.2.2)$$

i.e. in cylindrical coordinates ρ, φ, z the field is independent of z . At $\rho = \infty$ the field must assume its vacuum value, i.e. $|\Phi| = v$, in order to have finite energy. However, it may have a φ -dependent complex phase $\chi(\varphi)$

$$\Phi(\rho = \infty, \varphi) = v \exp(i\chi(\varphi)). \quad (10.2.3)$$

Points at infinity are parametrized by the polar angle φ , which is topologically a circle S^1 . In this case the vacuum manifold is also a circle, because $\mathcal{M} = U(1) = S^1$. It is parametrized by the angle χ . Hence, the field values at infinity $\Phi(\rho = \infty, \varphi)$ can be viewed as a mapping

$$\Phi : S^1 \rightarrow \mathcal{M}. \quad (10.2.4)$$

We know that such mappings are characterized by winding numbers in the homotopy group

$$\Pi_1[\mathcal{M}] = \Pi_1[S^1] = \mathbb{Z}. \quad (10.2.5)$$

Cosmic strings are field configurations with non-vanishing winding number. They are topological stable, since the winding cannot be eliminated by continuous deformations.

The simplest cosmic string has winding number 1. One can then use the identity map

$$\chi(\varphi) = \varphi. \quad (10.2.6)$$

Let us make the ansatz

$$\Phi(\rho, \varphi) = f(\rho) \exp(i\varphi), \quad f(\infty) = v. \quad (10.2.7)$$

At the z -axis at $\rho = 0$ the angle φ is not well-defined (coordinate singularity). For the field Φ to be well-defined, we must therefore demand $f(0) = 0$. Inserting this ansatz in the classical equations of motion

$$\partial_\mu \partial^\mu \Phi + m^2 \Phi + 2\lambda |\Phi|^2 \Phi = 0, \quad (10.2.8)$$

one obtains

$$\begin{aligned} \frac{1}{\rho} \frac{\partial}{\partial \rho} \left(\rho \frac{\partial \Phi}{\partial \rho} \right) + \frac{1}{\rho^2} \frac{\partial^2 \Phi}{\partial \rho^2} &= m^2 \Phi + 4\lambda |\Phi|^2 \Phi \Rightarrow \\ \frac{1}{\rho} \frac{\partial}{\partial \rho} (\rho f') \exp(i\varphi) - \frac{1}{\rho^2} f \exp(i\varphi) &= m^2 f \exp(i\varphi) + 2\lambda f^3 \exp(i\varphi) \Rightarrow \\ f'' + \frac{1}{\rho} f' - \frac{1}{\rho^2} f &= m^2 f + 2\lambda f^3. \end{aligned} \quad (10.2.9)$$

Although this equation cannot be solved analytically, it is easy to solve it numerically. The result is a monotonically rising function with $f(0) = 0$ and $f(\infty) = v = \sqrt{-m^2/2\lambda}$. The energy of the solution is concentrated around the z -axis, where $|\Phi|$ deviates from the vacuum value. The energy per unit length in the z -direction is known as the string tension μ . Due to its coupling to the massless Goldstone boson the tension of the string diverges logarithmically with its length.

The solution that we just discussed represents a so-called global cosmic string. Such a string cannot simply end. Either it is closed, or it extends through the whole Universe. When a $U(1)$ gauge symmetry gets spontaneously broken, string-like solutions still exist. They are then called local cosmic strings. Local strings interact with the massive gauge bosons that result when the massless Goldstone

boson is eaten. As a consequence, the infra-red logarithmic divergence of the string tension of global strings disappears. Indeed the string tension of a local cosmic string is finite.

The relevant homotopy group that decides about the existence of topologically stable cosmic string solutions is $\Pi_1[\mathcal{M}]$. In general, a group G may get spontaneously broken to a subgroup H . Then $\mathcal{M} = G/H$ and the corresponding homotopy group is

$$\Pi_1[\mathcal{M}] = \Pi_1[G/H]. \quad (10.2.10)$$

In our example we had $G = U(1)$ and $H = \{1\}$ such that

$$\Pi_1[\mathcal{M}] = \Pi_1[U(1)/\{1\}] = \Pi_1[U(1)] = \Pi_1[S^1] = \mathbb{Z}. \quad (10.2.11)$$

In the standard model, on the other hand, $G = SU(3) \otimes SU(2) \otimes U(1)$ and $H = SU(3) \otimes U(1)$, such that

$$\begin{aligned} \Pi_1[\mathcal{M}] &= \Pi_1[SU(3) \otimes SU(2) \otimes U(1)/SU(3) \otimes U(1)] = \\ &= \Pi_1[SU(2)] = \Pi_1[S^3] = \{0\}. \end{aligned} \quad (10.2.12)$$

Hence, there are no stable cosmic string solutions in the standard model. In some extensions of the standard model cosmic strings appear naturally, when a $U(1)$ symmetry breaks spontaneously. Then a network of cosmic strings emerges that extends throughout the whole Universe.

There is a general relation between homotopy groups. If $\Pi_n[G] = \{0\}$ then

$$\Pi_n[G/H] = \Pi_{n-1}[H]. \quad (10.2.13)$$

This implies that cosmic strings can also arise without explicit $U(1)$ factors. For example, if an $SU(2)$ symmetry would break spontaneously to $\mathbb{Z}(2)$, one would have

$$\Pi_1[SU(2)/\mathbb{Z}(2)] = \Pi_0[\mathbb{Z}(2)] = \mathbb{Z}(2). \quad (10.2.14)$$

In that case, two cosmic strings together could unwind each other, because there are only two topological sectors. Such strings are known as Alice strings.

Cosmic strings have interesting gravitational effects, because they curve space to a cone. The resulting deficit angle $\Delta\theta = 4G\mu$ leads to a gravitational lensing effect. This may eventually lead to their observation. Due to their gravitational effects, cosmic strings may act as seeds for galaxy formation. Just like the normal conducting state is restored inside an Abrikosov flux string in a superconductor, the vacuum can become superconducting inside a cosmic string. Such superconducting cosmic strings occur in models with two coupled $U(1)$ symmetries, one of

them the symmetry of electromagnetism, which is unbroken in the vacuum. The other $U(1)$ is broken and gives rise to the cosmic string. Inside the string the unbroken vacuum of this $U(1)$ is restored, but this may now lead to a breaking of electromagnetism. Superconducting cosmic strings can lead to large electric fields. They are dangerous objects, because they can explode. People have even tried to built time-machines using colliding cosmic strings. Unfortunately, in the examples studied so far the Universe always disappeared in a Big Crunch before the travel around a closed time-like curve could be completed. Clearly, cosmic strings are interesting objects for theoretical study. If they exist in our Universe is an open question, that can only be decided by observation.

10.3 Magnetic Monopoles

Magnetic monopoles are point-like topological excitations, which arise naturally in GUTs. A simplified (and hence unrealistic) GUT model is the $SU(2)$ Georgi-Glashow model. In the $SU(5)$ model we used a Higgs field in the adjoint representation to break the gauge symmetry spontaneously. We do the same in the simplified $SU(2)$ model, i.e.

$$\Phi(x) = \Phi^a(x)\sigma_a. \quad (10.3.1)$$

The Higgs field can be diagonalized by a unitary transformation (unitary gauge). The gauge fixed field is still invariant under $U(1)$ gauge transformations and hence the resulting symmetry breaking pattern is $SU(2) \rightarrow U(1)$. We identify this $U(1)$ with the gauge symmetry of electromagnetism. The Lagrange function is then given by

$$\mathcal{L}(\Phi, V_\mu) = \text{Tr} D^\mu \Phi D_\mu \Phi - V(\Phi) - \frac{1}{4e^2} \text{Tr} V^{\mu\nu} V_{\mu\nu}, \quad (10.3.2)$$

and again

$$V_{\mu\nu} = \partial_\mu V_\nu - \partial_\nu V_\mu + [V_\mu, V_\nu], \quad D_\mu \Phi = \partial_\mu \Phi + [V_\mu, \Phi]. \quad (10.3.3)$$

Let us look for spherically symmetric solutions of the field equations

$$\Phi(x) = \Phi(\vec{x}, t) = \Phi(r, \theta, \varphi). \quad (10.3.4)$$

At spatial infinity the scalar field must go to its vacuum value, i.e.

$$\text{Tr} \Phi(\infty, \theta, \varphi)^\dagger \Phi(\infty, \theta, \varphi) = v^2. \quad (10.3.5)$$

Spatial infinity is topologically a sphere S^2 . The vacuum manifold is thus given by

$$\mathcal{M} = SU(2)/U(1), \quad (10.3.6)$$

since a $U(1)$ symmetry remains after spontaneous symmetry breaking. The field configurations are thus characterized by winding numbers in the homotopy group

$$\Pi_2[SU(2)/U(1)] = \Pi_2[S^3/S^1]. \quad (10.3.7)$$

Since $\Pi_2[S^3] = \{0\}$ we obtain

$$\Pi_2[S^3/S^1] = \Pi_1[S^1] = \mathbb{Z}. \quad (10.3.8)$$

The integer winding number will turn out to be the magnetic charge. Again, we make an ansatz for the solution

$$\Phi^a(r, \theta, \varphi) = f(r) \frac{x^a}{r}, \quad A_i^a(r, \theta, \varphi) = g(r) \varepsilon_{iab} \frac{x^b}{r^2}, \quad (10.3.9)$$

and again $f(0) = 0$, $f(\infty) = v$, $g(0) = 0$, $g(\infty) = 1$. It is easy to solve the resulting equations for f and g numerically and one finds monotonically rising functions. In the unitary gauge $\Phi = \Phi^3 \sigma_3$ the 3-component of the field strength at large distances is given by

$$B_i^3 = \frac{1}{2} \varepsilon_{ijk} F_j k^3 = g \frac{r^i}{r^3}. \quad (10.3.10)$$

This is the field of a magnetic charge g . The magnetic charge is given by the Dirac quantization condition

$$g = \frac{2\pi}{e}. \quad (10.3.11)$$

The monopole mass turns out to be

$$M \approx gv, \quad (10.3.12)$$

i.e. GUT monopoles have a mass of about 10^{16} GeV — an elementary particle which is as heavy as a bacterium.

In the monopole core the scalar field goes to zero, i.e. the symmetric vacuum is restored inside the monopole. Effectively, one is then back in the high temperature phase of the GUT theory. In this phase baryon number violating processes are un-suppressed. Hence, magnetic monopoles can catalyze proton decay.

In the standard model there are no magnetic monopoles, because

$$\Pi_2[SU(3) \otimes SU(2) \otimes U(1)/SU(3) \otimes U(1)] = \Pi_2[SU(2)] = \Pi_2[S^3] = \{0\}. \quad (10.3.13)$$

In the $SU(5)$ GUT, on the other hand,

$$\begin{aligned}\Pi_2[SU(5)/SU(3) \otimes SU(2) \otimes U(1)] &= \Pi_1[SU(3) \otimes SU(2) \otimes U(1)] = \\ \Pi_1[U(1)] &= \Pi_1[S^1] = \mathbb{Z}.\end{aligned}\tag{10.3.14}$$

The same is true for other GUT theories like $SO(10)$ or E_6 , since then again $\Pi_2[G] = 0$, and the $U(1)$ symmetry of the standard model remains as an unbroken subgroup. Hence, magnetic monopoles are unavoidable in GUTs.

10.4 The Kibble Mechanism

Why have no monopoles been detected yet? Maxwell's theory with $\vec{\nabla} \cdot \vec{B} = 0$ agrees very well with experiments. Would it be possible that GUTs have monopole solutions, but monopoles have simply not been created? The Kibble mechanism shows that this is impossible. One assumes that in the very early Universe a GUT phase transition has taken place at which the $SU(5)$ symmetry has been broken spontaneously to $SU(3) \otimes SU(2) \otimes U(1)$. In the high temperature phase the scalar field has a vanishing vacuum expectation value, which means that the field points in random directions, which add up to zero when integrated over small spatial regions. The size of such regions is given by a finite correlation length. This length can at most be the horizon size

$$d_H = R(t) \int_0^t dt' \frac{1}{R(t')} = 2t.\tag{10.4.1}$$

The above formula applies to a radiation dominated Universe. When the system undergoes the phase transition and enters the low-temperature phase the field fluctuations are frozen, and Φ gets its vacuum expectation value. The originally different random orientations of the scalar field then unavoidably lead to the creation of topological excitations, which depending on the symmetry breaking pattern could be monopoles, cosmic strings or domain walls. In GUTs it is natural to assume that at least one monopole is created per horizon volume at the time of the phase transition. One has $d_H \approx m_P/T^2$ implying a monopole density

$$n_M \approx d_H^{-3} \approx \frac{T_c^6}{m_P^3}.\tag{10.4.2}$$

The entropy density is proportional to T^3 and we then obtain

$$\frac{n_M}{s} \approx (T_c/m_P)^3 \approx (10^{14}\text{GeV}/10^{19}\text{GeV})^3 \approx 10^{-13}.\tag{10.4.3}$$

Since both the entropy and the monopole number are conserved during the expansion of the Universe, this number remains unchanged. Today the ratio of the baryon density and entropy is $n_B/s \approx 10^{-9}$ and baryonic matter contributes about 0.1 to Ω . Since a monopole weighs 10^{16} times a baryon, the monopoles contribute about 10^{11} to Ω , i.e. they totally dominate the energy density and are completely inconsistent with cosmological observations.

If monopoles were present in the Universe, they would accumulate in neutron stars and catalyze proton decay. This would lead to very strong emission of radiation, which is not observed. Hence, the question arises where the GUT monopoles went. A possible scenario uses confinement by cosmic strings. One then assumes that the Universe has been superconducting for a while. Monopoles and anti-monopoles are then connected by Abrikosov cosmic strings. The strings cost energy proportional to their length, and thus pull monopoles and anti-monopoles together until they annihilate. Inflation offers another way to get rid of the unwanted monopoles, which will be discussed later.

Chapter 11

Axions

The topological properties of the gluon field give rise to several problems in the standard model. One is the strong CP problem related to the presence of the θ -vacuum angle. If θ were non-zero it would give rise to an electric dipole moment of the neutron. Since the actual dipole moment is unmeasurably small, one concludes that $|\theta| < 10^{-9}$, a very small number that calls for an explanation. Of course, in the context of QCD alone one could postulate CP symmetry, and thus force $\theta = 0$ (or $\theta = \pi$). When QCD is embedded in the standard model, this does not work any longer, because CP is explicitly violated by the complex phase of the Kobayashi-Maskawa matrix. Thus, $\theta = 0$ is not protected by CP symmetry any more and even if one puts $\theta = 0$ at the classical level, it would be regenerated by renormalization effects. Another way out would be to assume that gluon field configurations with non-vanishing topological charge are negligible in the QCD path integral. This, however, also does not work because there is also the so-called $U(1)$ -problem in QCD. The problem is to explain why the η' -meson has a large mass and hence is not a Goldstone boson. This is qualitatively understood based on the Adler-Bell-Jackiw anomaly — the axial $U(1)$ symmetry of QCD is simply explicitly broken. To solve the $U(1)$ -problem quantitatively — i.e. to explain the large value of the η' -mass — requires gluon field configurations with non-zero topological charge to appear frequently in the path integral. This is confirmed by lattice calculations and indeed offers a nice explanation of the $U(1)$ -problem. However, if we use topologically non-trivial configurations to solve the $U(1)$ -problem, we cannot ignore these configurations when we face the strong CP-problem.

Peccei and Quinn have suggested an extension of the standard model that indeed solves the strong CP-problem. In their model there are two Higgs doublets,

rather than one as in the standard model. This situation also naturally arises in supersymmetric extensions of the standard model. As a consequence of the presence of the second Higgs field there is an extra $U(1)_{PQ}$ — a so-called Peccei-Quinn symmetry in the problem. When this symmetry breaks spontaneously, it dynamically favors a vacuum in which $\theta = 0$. As usual in the case of a spontaneously broken global symmetry, a massless Goldstone boson arises. This Goldstone boson is known as the axion. Still, the axion is not exactly massless, because the $U(1)_{PQ}$ is also explicitly broken by the anomaly. As a consequence, the axion picks up a small mass. Unfortunately, nobody has ever detected an axion despite numerous experimental efforts and it is unclear if this is the right solution of the strong CP problem. Although the original Peccei-Quinn model was soon ruled out by experiments, the symmetry breaking scale of the model can be shifted to higher energy scales making the axion invisible.

Axions are very interesting players in the Universe. They couple only weakly to ordinary matter, but they still have interesting effects. First of all, they are massive and could provide enough energy to close the Universe and make $\Omega = 1$. If it exists the axion can also shorten the life-time of stars. Stars live so long, because they cannot get rid of their energy by radiation very fast. For example, a photon that is generated in a nuclear reaction in the center of the sun spends 10^7 years before it reaches the sun's surface, simply because its electromagnetic cross section with the charged matter in the sun is large. An axion, on the other hand, interacts weakly and can thus get out much faster. Like neutrinos, axions can therefore act as a super coolant for stars. The observed life-time of stars can thus be used to put astro-physical limits on axion parameters like the axion mass. Axions can be generated in the early Universe in multiple ways. First, they can simply be thermally produced. Then they can be generated by a disoriented $U(1)_{PQ}$ condensate. This mechanism is similar to the recently discussed pion production via a disoriented chiral condensate in a heavy ion collision generating a quark-gluon plasma. Also, as we learned in the previous chapter, the spontaneous breakdown of a $U(1)$ symmetry is accompanied by the generation of cosmic strings. Indeed, if the axion exists, axionic cosmic strings should exist as well. A network of such fluctuating strings could radiate energy by emitting the corresponding Goldstone bosons, namely axions. Unlike other global strings, axionic strings have a finite string tension, because the axion itself is not a massless Goldstone boson, but picks up a small mass from the anomaly. Before we discuss the axion in more detail, let us first discuss the solution of the $U(1)$ -problem, which suggests that topologically non-trivial gluon field configurations indeed exist.

11.1 The $U(1)$ -Problem

The chiral symmetry of the classical QCD Lagrange function is $U(N_f)_L \otimes U(N_f)_R$, while in the spectrum only the $SU(N_f)_{L+R} \otimes U(1)_{L=R} = U(N_f)_{L=R}$ symmetries are manifest. According to the Goldstone theorem one might hence expect $N_f^2 + N_f^2 - N_f^2 = N_f^2$ Goldstone bosons, while in fact one finds only $N_f^2 - 1$ Goldstone bosons in QCD. The missing Goldstone boson should be a pseudoscalar, flavor-scalar particle. The lightest particle with these quantum numbers is the η' -meson. However, its mass is $M_{\eta'} = 0.958$ GeV, which is far too heavy for a Goldstone boson. The question why the η' -meson is so heavy is the so-called $U(1)$ -problem of QCD. At the end the question is why the axial $U(1)$ symmetry is not spontaneously broken, although it is also not manifest in the spectrum. It took a while before people realized that the axial $U(1)$ is not really a symmetry of QCD. Although the symmetry is present in the classical Lagrange density, it cannot be maintained under quantization because it has an anomaly. This explains qualitatively why the η' -meson is not a Goldstone boson. To understand the problem more quantitatively, one must consider the origin of the quantum mechanical symmetry breaking in more detail. It turns out that topologically non-trivial configurations of the gluon field — for example instantons — give mass to the η' -meson. If the color symmetry would be $SU(N_c)$ instead of $SU(3)$, the explicit axial $U(1)$ breaking via the anomaly would disappear in the large N_c limit. In this limit the η' -meson does indeed become a Goldstone boson. For large but finite N_c the η' -meson gets a mass proportional to the topological susceptibility — the vacuum value of the topological charge squared per space-time volume — evaluated in the pure glue theory.

Qualitatively one understands why the η' -meson is not a Goldstone boson, because the axial $U(1)$ -symmetry is explicitly broken by the Adler-Bell-Jackiw anomaly

$$\partial_\mu J_\mu^5(x) = 2N_f P(x), \quad (11.1.1)$$

where P is the Chern-Pontryagin density. However, the question arises how strong this breaking really is, and how it affects the η' -mass quantitatively. To understand this issue we consider QCD with a large number of colors, i.e. we replace the gauge group $SU(3)$ by $SU(N_c)$.

It is interesting that large N_c QCD is simpler than real QCD, but still it is too complicated to solve it analytically. Still, one can classify the subset of Feynman diagrams that contribute in the large N_c limit. An essential observation is that for many colors the distinction between $SU(N_c)$ and $U(N_c)$ becomes irrelevant. Then each gluon propagator in a Feynman diagram may be replaced formally by

the color flow of a quark-antiquark pair. In this way any large N_c QCD diagram can be represented as a quark diagram. For the gluon self energy diagram, for example, one finds an internal quark loop which yields a color factor N_c and each vertex gives a factor g_s , such that the diagram diverges as $g_s^2 N_c$. We absorb this divergence in a redefinition of the coupling constant by defining

$$g^2 = g_s^2 N_c, \quad (11.1.2)$$

and we perform the large N_c limit such that g_s goes to zero but g remains finite. Let us now consider a planar 2-loop diagram contributing to the gluon self energy. There are two internal loops and hence there is a factor N_c^2 . Also there are four vertices contributing factors $g_s^4 = g^4/N_c^2$ and the whole diagram is proportional to g^4 and hence it is finite. Let us also consider a planar 4-loop diagram. It has a factor N_c^4 together with six 3-gluon vertices that give a factor $g_s^6 = g^6/N_c^3$ and a 4-gluon vertex that gives a factor $g_s^2 = g^2/N_c$. Altogether the diagram is proportional to g^8 and again it is finite as N_c goes to infinity. Next let us consider a non-planar 4-loop diagram. The color flow is such that now there is only one color factor N_c but there is a factor $g_s^6 = g^6/N_c^3$ from the vertices. Hence the total factor is g^6/N_c^2 which vanishes in the large N_c limit. In general any non-planar gluon diagram vanishes in the large N_c limit. Planar diagrams, on the other hand, survive in the limit. In particular, if we add another propagator to a planar diagram such that it remains planar, we add two 3-gluon vertices and hence a factor $g_s^2 = g^2/N_c$, and we cut an existing loop into two pieces, thus introducing an extra loop color factor N_c . The total weight remains of order 1. Now consider the quark contribution to the gluon propagator. There is no color factor N_c for this diagram, and still there are two quark-gluon vertices contributing a factor $g_s^2 = g^2/N_c$. Hence this diagram disappears in the large N_c limit. Similarly, all diagrams with internal quark loops vanish at large N_c . Even though this eliminates a huge class of diagrams, the remaining planar gluon diagrams are still too complicated to be summed up analytically. Still, the above N_c counting allows us to understand some aspects of the QCD dynamics.

In the large N_c limit QCD reduces to a theory of mesons and glueballs, while the baryons disappear. This can be understood in the constituent quark model. In $SU(N_c)$ a color singlet baryon consists of N_c quarks, each contributing the constituent quark mass to the total baryon mass. Hence the baryon mass is proportional to N_c such that baryons are infinitely heavy (and hence disappear) in the large N_c limit. Mesons, on the other hand, still consist of a quark and an anti-quark, such that their mass remains finite.

Also the topology of the gluon field is affected in the large N_c limit. We have

derived the instanton action bound

$$S[G_\mu] \geq \frac{8\pi^2}{g_s^2} |Q[G_\mu]| = \frac{8\pi^2 N_c}{g^2} |Q[G_\mu]|, \quad (11.1.3)$$

which is valid for all $SU(N_c)$. In the large N_c limit the action of an instanton diverges, and topologically non-trivial field configurations are eliminated from the Feynman path integral. This means that the source of quantum mechanical symmetry breaking via the anomaly disappears, and the η' -meson should indeed become a Goldstone boson in the large N_c limit. In that case one should be able to derive a mass formula for the η' -meson just like for the Goldstone pion. The pion mass resulted from an explicit chiral symmetry breaking due to a finite quark mass. Similarly, the η' -mass results from an explicit axial $U(1)$ breaking via the anomaly due to a finite N_c . This can be computed as a $1/N_c$ effect.

Let us consider the so-called topological susceptibility as the integrated correlation function of two Chern-Pontryagin densities

$$\chi_t = \int d^4x \, {}_{pg} \langle 0 | P(0) P(x) | 0 \rangle_{pg} \frac{Q^2}{V} \quad (11.1.4)$$

in the pure gluon theory (without quarks). Here $|0\rangle_{pg}$ is the vacuum of the pure gluon theory, and V is the volume of space-time. When we add massless quarks the Atiyah-Singer index theorem implies that the topological charge — and hence χ_t — vanishes, because the zero modes of the Dirac operator eliminate topologically nontrivial field configurations. Therefore in full QCD (with massless quarks)

$$\int d^4x \, \langle 0 | P(0) P(x) | 0 \rangle = 0, \quad (11.1.5)$$

where $|0\rangle$ is the full QCD vacuum. In the large N_c limit the effects of quarks are $1/N_c$ suppressed. Therefore it is unclear how they can eliminate the topological susceptibility of the pure gluon theory. In the large N_c limit the quark effects manifest themselves entirely in terms of mesons. One finds

$$\chi_t - \sum_m \frac{\langle 0 | P | m \rangle \langle m | P | 0 \rangle}{M_m^2} = 0, \quad (11.1.6)$$

where the sum runs over all meson states and M_m are the corresponding meson masses. Using large N_c techniques one can show that $|\langle 0 | P | m \rangle|^2$ is of order $1/N_c$, while χ_t is of order 1. If also all meson masses would be of order 1 there would be a contradiction. The puzzle gets resolved when one assumes that the lightest flavorscalar, pseudoscalar meson — the η' — has in fact a mass of order $1/N_c$, such that

$$\chi_t = \frac{|\langle 0 | P | \eta' \rangle|^2}{M_{\eta'}^2}. \quad (11.1.7)$$

Using the anomaly equation one obtains

$$\langle 0|P|\eta'\rangle = \frac{1}{2N_f}\langle 0|\partial_\mu A_\mu|\eta'\rangle = \frac{1}{\sqrt{2N_f}}M_{\eta'}^2 f_{\eta'}. \quad (11.1.8)$$

In the large N_c limit $f_{\eta'} = f_\pi$ and we arrive at the Witten-Veneziano formula

$$\chi_t = \frac{f_\pi^2 M_{\eta'}^2}{2N_f}. \quad (11.1.9)$$

In this equation χ_t is of order 1, f_π^2 is of order N_c and $M_{\eta'}^2$ is of order $1/N_c$. This means that the η' -meson is indeed a Goldstone boson in a world with infinitely many colors. At finite N_c the anomaly arises leading to an explicit axial $U(1)$ symmetry breaking proportional to $1/N_c$. The pseudo Goldstone boson mass squared is hence proportional to $1/N_c$. So far we have assumed that all quarks are massless. When a nonzero s quark mass is taken into account the formula changes to

$$\chi_t = \frac{1}{6}f_\pi^2(M_{\eta'}^2 + M_\eta^2 - 2M_K^2) = (0.180\text{GeV})^4. \quad (11.1.10)$$

Lattice calculations are at least roughly consistent with this value, which supports this solution of the $U(1)$ -problem.

11.2 A Solution of the Strong CP Problem

At the level of QCD alone the strong CP-problem is to understand why the term θQ does not occur in the QCD Lagrangian. Since this is the only term that breaks CP, one can always postulate that CP is conserved and “solve” the problem that way. When QCD is embedded in the standard model, this is no longer possible, because CP is then explicitly broken already by the complex phase of the Kobayashi-Maskawa matrix. Let us discuss this matrix first. The scalar field Φ in the standard model is a complex $SU(2)$ doublet, i.e. under the $SU(3) \otimes SU(2) \otimes U(1)$ gauge group it transforms as $\{1, 2\}_{-1/2}$. Due to the special properties of $SU(2)$ one can use Φ to construct another field

$$\tilde{\Phi} = i\sigma_2 \Phi^*, \quad (11.2.1)$$

which transforms as $\{1, 2\}_{1/2}$. The three generations of quarks can be written as

$$\begin{aligned} q_{Li} &= \left\{ \begin{pmatrix} u_L \\ d_L \end{pmatrix}, \begin{pmatrix} c_L \\ s_L \end{pmatrix}, \begin{pmatrix} t_L \\ b_L \end{pmatrix} \right\}, \\ u_{Ri} &= \{u_R, c_R, t_R\}, \quad d_{Ri} = \{d_R, s_R, b_R\}. \end{aligned} \quad (11.2.2)$$

The Yukawa couplings that give rise to the quark masses are then given by the Lagrange density

$$\mathcal{L}(\Phi, q_L, u_R, d_R) = -f_{ij}^u \bar{q}_{Li} \Phi u_{Rj} + f_{ij}^d \bar{q}_{Li} \tilde{\Phi} d_{Ri} + \text{h.c.} \quad (11.2.3)$$

After spontaneous symmetry breaking the scalar field gets its vacuum value

$$\langle \Phi \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} v \\ 0 \end{pmatrix}, \quad \langle \tilde{\Phi} \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}. \quad (11.2.4)$$

Since the Yukawa couplings f_{ij}^u and f_{ij}^d are not diagonal in the generation indices i and j one obtains mass matrices

$$M_i^u j = \frac{1}{\sqrt{2}} f_{ij}^u v, \quad M_i^d j = \frac{1}{\sqrt{2}} f_{ij}^d v. \quad (11.2.5)$$

The quark masses are the eigenvalues of these matrices, which can be diagonalized by bi-unitary transformations

$$(U_L^u)^\dagger M^u U_R^u = \text{diag}(m_u, m_c, m_t), \quad (U_L^d)^\dagger M^d U_R^d = \text{diag}(m_d, m_s, m_b). \quad (11.2.6)$$

Of course, the bi-unitary transformation must be applied consistently to the rest of the Lagrangian, which leads to mixing between mass eigenstates described by the Kobayashi-Maskawa matrix

$$V = (U_L^u)^\dagger U_L^d. \quad (11.2.7)$$

To diagonalize the quark mass matrices we have used the bi-unitary transformations consisting of left and right-handed pieces. In general, the transformations will involve axial $U(1)$ transformations, which are not a symmetry of the quantum theory due to the Adler-Bell-Jackiw anomaly. As a consequence of the anomaly, the bi-unitary transformations will therefore induce a topological term in the QCD Lagrangian even if such a term was not put in at the classical level. The induced θ -term is proportional to the phase of the quark mass matrix, such that in total

$$\bar{\theta} = \theta + \arg \det M \quad (11.2.8)$$

is the relevant vacuum angle. It is indeed $\bar{\theta}$ to which the experimental bound from the electric dipole moment of the neutron applies.

The idea of Peccei and Quinn is to rotate $\bar{\theta}$ to zero by enlarging the symmetry of the standard model by an additional $U(1)_{PQ}$. Let us again consider the Yukawa couplings in the standard model of eq.(11.2.3). They are invariant under axial $U(1)$ transformations

$$\begin{pmatrix} u'_L \\ d'_L \end{pmatrix} = \exp(-i\alpha) \begin{pmatrix} u_L \\ d_L \end{pmatrix}, \quad u'_R = \exp(i\alpha) u_R, \quad d'_R = \exp(i\alpha) d_R, \quad (11.2.9)$$

only if

$$\Phi' = \exp(-2i\alpha)\Phi, \quad \tilde{\Phi}' = \exp(-2i\alpha)\tilde{\Phi}. \quad (11.2.10)$$

This, however, is inconsistent with $\tilde{\Phi} = i\sigma_2\Phi^*$. Hence, the standard model Lagrangian is indeed not invariant under axial $U(1)$ transformations. Peccei and Quinn suggested to enlarge the symmetry of the standard model by introducing a second Higgs doublet that transforms like $\tilde{\Phi}$ but that is unrelated to the original Higgs field Φ . Then the system has an extra $U(1)$ symmetry, which can be used to rotate away the unwanted θ term. Of course, one still wants to make use of the Higgs mechanism and break the $SU(2)_L \times U(1)_Y$ gauge symmetry spontaneously to $U(1)_{em}$. Then the two Higgs fields require vacuum expectation values, which also break the additional $U(1)_{PQ}$ symmetry. As a consequence, a Goldstone boson arises, which is known as the axion.

11.3 Axion Properties

The scalar potential $V(\Phi, \tilde{\Phi})$ in the extension of the standard model is invariant against $SU(2)_L \otimes U(1)_Y \otimes U(1)_{PQ}$ transformations. When this symmetry breaks down spontaneously to $U(1)_{em}$, there are $3 + 1 + 1 - 1 = 4$ massless Goldstone bosons. Three are eaten by the gauge bosons and become the longitudinal components of Z_0 and W^\pm . Since $U(1)_{PQ}$ is a global symmetry, the fourth Goldstone boson remains uneaten. The neutral components of the massless fields can be written as

$$\Phi^0 = \frac{1}{\sqrt{2}}(v + \rho + i\sigma), \quad \tilde{\Phi}^0 = \frac{1}{\sqrt{2}}(\tilde{v} + \tilde{\rho} + i\tilde{\sigma}). \quad (11.3.1)$$

Since Φ^0 and $\tilde{\Phi}^0$ have opposite hyper-charge the neutral combination that is eaten by Z^0 is

$$\sigma_{Z^0} = \frac{v\sigma - \tilde{v}\tilde{\sigma}}{\sqrt{v^2 + \tilde{v}^2}}, \quad (11.3.2)$$

while the orthogonal combination

$$\sigma_a = \frac{\tilde{v}\sigma + v\tilde{\sigma}}{\sqrt{v^2 + \tilde{v}^2}}, \quad (11.3.3)$$

is the axion. In the model with two Higgs fields the combination $\sqrt{v^2 + \tilde{v}^2}$ represents the electroweak scale. The remaining combination $x = \tilde{v}/v$ is a free parameter of axion models. It is important to note that the axion is not exactly massless. This is a consequence of the axial anomaly, which explicitly breaks the $U(1)_{PQ}$ symmetry. In the chiral limit this breaking would disappear, because θ

could then be rotated away even without the Peccei-Quinn symmetry. One can show that the induced axion mass is given by

$$M_a = \frac{M_\pi f_\pi}{\sqrt{v^2 + \tilde{v}^2}} \frac{N_f}{2} \left(x + \frac{1}{x}\right) \frac{\sqrt{m_u m_d}}{m_u + m_d}. \quad (11.3.4)$$

Here M_π is the pion mass, f_π is the pion decay constant, and N_f is the number of flavors. Inserting the values for M_π , f_π and the electroweak scale, one obtains

$$M_a = 12.5 N_f \left(x + \frac{1}{x}\right) \text{keV}. \quad (11.3.5)$$

For $x \approx 1$ and $N_f = 6$ one has $M_a > 150$ keV, which makes the axion lighter than an e^+e^- pair. It can still decay into two photons, and thus is not absolutely stable, but would be rather long lived. If x is such that the e^+e^- channel opens up, the live time would be much shorter. The so-called standard axion with $x \approx 1$ has been ruled out experimentally. Still, the possibility remains that $U(1)_{PQ}$ is broken at a scale high above the electroweak scale. Such an invisible axion would be very light, it would couple extremely weakly to ordinary matter, and it would have intriguing astrophysical and cosmological implications.

11.4 Cosmological Implications of the Axion

Invisible axions are light and interact only weakly, because the axion coupling constants are proportional to the mass. Still, unless they are too light — and thus too weakly interacting — axions can cool stars very efficiently. Their low interaction cross section allows them to carry away energy more easily than the more strongly coupled photon. A sufficiently interacting axion could shorten the life-time of stars by a substantial amount. From the observed life-time one can hence infer an upper limit on the axion mass. In this way invisible axions heavier than 1 eV have been ruled out. This implies that the Peccei-Quinn symmetry breaking scale must be above 10^7 GeV. If they exist, axions would also affect the cooling of a neutron star that forms after a supernova explosion. There would be less energy taken away by neutrinos. The observed neutrino burst of the supernova SN 1987A would have consisted of fewer neutrinos if axion had also cooled the neutron star. This astrophysical observation excludes axions of masses between 10^{-3} and 0.02 MeV — a range that cannot be investigated in the laboratory.

There are various mechanisms in the early Universe that can lead to the generation of axions. The simplest is via thermal excitation. One can estimate that

thermally generated axions must be rather heavy in order to contribute substantially to the energy density of the Universe. In fact, thermal axions cannot close the Universe, because the required mass is already ruled out by the astrophysical limits. Another interesting mechanism for axion production is related to the fact that the axial anomaly breaks $U(1)_{PQ}$. As a consequence, the axion picks up a small mass, and the scalar potential of the two Higgs fields then has a unique minimum. This happens at typical QCD energy scales through instanton effects similar to the ones that give mass to the η' -meson. At higher energies the axion is effectively massless, and the scalar potential has several degenerate minima. Hence, there are several degenerate vacua labelled by the θ -parameter, and any value of θ is equally probable. Then different regions of the Universe must have been in different θ -vacua. When the anomaly breaks the symmetry, $\theta = 0$ is singled out as the unique minimum. In order to minimize its energy, the scalar field then “rolls” down to this minimum, and oscillates about it. The oscillations are damped by axion emission, and finally $\theta = 0$ is reached everywhere in the Universe. The axions produced in this way would form a Bose condensate that could close the Universe for an axion mass in the 10^{-5} eV range. This makes the axion an attractive candidate for the missing dark matter in the Universe. The axion production mechanism via a disoriented Peccei-Quinn condensate is very similar to the pion-production mechanism that has recently been discussed via disorienting the chiral condensate in a heavy ion collision. At temperatures high above the QCD scale $U(1)_{PQ}$ is almost an exact global symmetry, which gets spontaneously broken at some high scale. As we have seen before, this necessarily leads to the generation of a network of cosmic strings. Such a string network can lower its energy by radiating axions. Once the axion mass becomes important, the string solutions become unstable, and the string network disappears, again leading to axion emission. This production mechanism may also lead to enough axions to close the Universe.

Chapter 12

Inflation

The standard model of cosmology, i.e. a Friedmann Universe with Robertson-Walker metric, that is first radiation and then matter dominated, is very successful in describing phenomena in the early Universe. The synthesis of light nuclei at about 1 sec to 1 min after the Big Bang provides the earliest test of the standard model of cosmology. Also the explanation of the cosmic background radiation should be viewed as a big success. We have seen that the standard model of cosmology together with the standard model of particle physics and its extensions (e.g. GUT) lead to interesting phenomena in the extremely early Universe, like phase transitions in the strong and electroweak sector, or the generation of magnetic monopoles at a GUT phase transition. Although the standard model of particle physics agrees very well with experiments, from a theoretical point of view it is not completely satisfactory due to its large number of parameters. Also the standard model of cosmology has some problems. On the one hand, we know that due to the classical treatment of gravity it cannot be valid before the Planck time. On the other hand, it contains several parameters in the form of initial conditions, which must be fine tuned in order to explain the large age of our Universe. Such fine tuning appears unnatural to most cosmologists. The Robertson-Walker ansatz for the metric was motivated by observation (e.g. by the observed isotropy of the cosmic background radiation). In the framework of the standard model, it remains an open question why our Universe is isotropic and homogeneous on such large scales. On the other hand, on smaller (but still very large) scales we observe a lot of structure, like galaxies, galaxy clusters, super-clusters as well as large voids. In the framework of the standard model it is unclear how these deviations from homogeneity have evolved from small initial fluctuations. Finally, GUTs — which can at least qualitatively account for

the baryon asymmetry — unavoidably lead to monopole creation via the Kibble mechanism, but monopoles have not been observed. All these puzzles find an answer by Alan Guth's idea of the inflationary Universe.

One then assumes that e.g. at a first order GUT phase transition a scalar field remained in the (e.g. $SU(5)$) symmetric phase, i.e. at the unstable minimum of the temperature-dependent effective potential, while the true minimum corresponds to the broken phase. The energy of the scalar field then acts as a cosmological constant, which leads to an exponential, inflationary expansion of the Universe. In this way any effect of earlier initial conditions is eliminated, and the Universe becomes homogeneous and isotropic on the largest scales. Eventually present magnetic monopoles get extremely diluted, and are effectively eliminated from the Universe. Furthermore, the quantum fluctuations of the scalar field give rise to initial inhomogeneities, which may have evolved into the structures observed today. The exponential expansion leads to a tremendous increase of the scale factor, and hence to a flat space. This implies $\Omega = 1$, and hence requires a sufficient amount of dark matter.

When $\Omega = 1$, this also explains the large age of our Universe (compared to the Planck time) and hence an important condition for our own existence. Hence, the inflationary Universe allows us to avoid the anthropic principle, which explains the large age of the Universe by the mere fact that we exist: we could not have developed in a short-lived Universe. Alan Guth's original idea has been modified by Linde who has introduced the scenario of chaotic inflation. One then assumes that initially the scalar field assumes some chaotic random values, and hence only certain regions of space undergo inflation. In this way different parts of the Universe with possibly different vacuum structure evolve. It should then be viewed as a historical fact, that in our local vacuum bubble the symmetry was broken to $SU(3) \otimes SU(2) \otimes U(1)$ and not to something else. Linde still uses the anthropic principle to explain why this is so: in another vacuum we could not have evolved.

Since at present the Universe does no longer expand exponentially, inflation (if it ever took place) must have come to an end. This must have been associated with the scalar field “rolling” down to the stable minimum of broken symmetry. In this process an enormous amount of latent heat is released, which manifests itself in an enormous entropy generation. At present that entropy is in the cosmic background radiation. The idea of inflation thus naturally leads to a Universe similar to the one we live in: old, flat and homogeneous and isotropic. Still, inflation should not be viewed as a well-established fact. It is a very attractive idea that still needs to be formulated in a realistic particle physics framework, and it

needs to be tested in observations. A strong indication for an inflationary epoch would be the determination of $\Omega = 1$. If our Universe turns out to be flat and eternally expanding, the idea of inflation will become part of an extended standard model of cosmology. Still, this would not solve all problems of cosmology. In particular, the question of the cosmological constant would remain unsolved: why is the vacuum energy so small today, such that the Universe no longer expands exponentially? To answer this question one probably needs a quantum theory of gravity, one of the great challenges in theoretical physics.

12.1 Deficiencies of the Standard Cosmology

Although the standard model of cosmology is very successful in many respects, it leaves some important questions unanswered. First, there is the question of the age of the Universe and, related to that, its flatness, which is also connected with the enormous entropy in the cosmic background radiation. To understand the age problem, let us consider the Friedmann equation

$$\frac{\dot{R}^2}{R^2} + \frac{k}{R^2} = \frac{8\pi}{3}G\rho \Rightarrow \frac{k}{H^2 R^2} = \frac{8\pi G\rho}{3H^2} - 1 = \frac{\rho}{\rho_c} - 1 = \Omega - 1, \quad (12.1.1)$$

where $H = \dot{R}/R$ is the Hubble parameter, and

$$\rho_c = \frac{3H^2}{8\pi G} \quad (12.1.2)$$

is the critical density. It is an observational fact, that our Universe has an energy density close to the critical one, i.e.

$$\Omega \in [0.1, 4]. \quad (12.1.3)$$

What does this mean for earlier times? As long as the Universe was matter dominated, one has

$$R(t) \propto t^{2/3} \Rightarrow H(t) = \frac{\dot{R}(t)}{R(t)} = \frac{2}{3t}, \quad (12.1.4)$$

and thus

$$\Omega(t) - 1 = \frac{k}{H(t)^2 R(t)^2} \propto \frac{t^2}{t^{4/3}} = t^{2/3}. \quad (12.1.5)$$

The Universe started to be matter dominated at an age of 10^5 years, and today it is about 10^{10} years old. Hence, when the cosmic background radiation decoupled, we had

$$\Omega(t) - 1 = \left(\frac{10^5}{10^{10}}\right)^{2/3} \approx 10^{-3}, \quad (12.1.6)$$

i.e. the density was even closer to the critical one than it is today. At earlier times the Universe was radiation dominated with

$$R(t) \propto t^{1/2} \Rightarrow H(t) = \frac{\dot{R}(t)}{R(t)} = \frac{1}{2t}, \quad (12.1.7)$$

and hence

$$\Omega(t) - 1 = \frac{k}{H(t)^2 R(t)^2} \propto \frac{t^2}{t} = t. \quad (12.1.8)$$

At the time when nucleosynthesis started, i.e. about 1 sec after the bang, one finds

$$\Omega(t) - 1 = 10^{-3} \frac{10^{-7}}{10^5} = 10^{-15}, \quad (12.1.9)$$

which again means that the density was extremely close to critical. Going further back in time, e.g. to a GUT phase transition at about 10^{-34} sec after the bang, one obtains

$$\Omega(t) - 1 = 10^{-15} 10^{-34} = 10^{-49}, \quad (12.1.10)$$

and if one goes back to the Planck time, 10^{-44} sec, one even finds

$$\Omega(t) - 1 = 10^{-15} 10^{-44} = 10^{-59}. \quad (12.1.11)$$

At first sight, today a value of Ω close to one seems not unnatural, but who has fine-tuned the energy density to sixty decimal places when the Universe was created at the Planck time?

This problem can also be expressed in terms of the entropy. In the standard cosmology the total entropy S in a comoving volume of size $R(t)^3$ is conserved. In the radiation dominated epoch the energy density is

$$\rho = \frac{\pi^2}{30} g_* T^4, \quad (12.1.12)$$

and the entropy density is

$$s = \frac{2\pi^2}{45} g_* T^3 = \frac{S}{R^3}. \quad (12.1.13)$$

At the Planck time the temperature is equal to the Planck mass, i.e. $T = m_P \approx 10^{19}$ GeV, such that then

$$\begin{aligned} \frac{\Omega - 1}{\Omega} &= \frac{k}{H^2 R^2} \frac{3H^2}{8\pi G \rho} = \frac{3k}{8\pi G \rho R^2} = \frac{3k}{8\pi G \rho} \left(\frac{s}{S}\right)^{2/3} \\ &= \frac{3k}{8\pi} \frac{30}{\pi^2 g_*} \left(\frac{2\pi^2}{45} g_*\right)^{2/3} \frac{1}{G m_P^2} S^{-2/3} \approx 0.1 S^{-2/3}. \end{aligned} \quad (12.1.14)$$

Using $\Omega - 1 \approx 10^{-59}$ one obtains the enormous entropy $S \approx 10^{87}$ per comoving volume. Today this entropy is contained in the cosmic background radiation.

Not only the enormous entropy of the cosmic background radiation cannot be understood in the framework of the standard cosmology. Also its isotropy presents a puzzle. The observed isotropy of the cosmic background radiation motivated the Robertson-Walker ansatz for the metric, but one does not understand where the isotropy comes from. In a radiation or matter dominated Robertson-Walker space-time the distance to the horizon

$$d_H(t) = R(t) \int_0^t \frac{dt'}{R(t')} \quad (12.1.15)$$

is finite. In a matter dominated Universe, e.g. one has $d_H(t) \propto 3t$. This is the maximal distance between two events that are causally connected. When matter and radiation decoupled at about $t_d = 10^5$ years after the bang, physical processes could establish thermal equilibrium over scales of about

$$d_H(t_d) \propto 3t_d. \quad (12.1.16)$$

Today, photons of the cosmic background radiation reach us from different directions. Their regions of generation at time t_d cannot have been in causal contact, and still the photons have almost exactly the same temperature.

Let us consider a photon that was emitted at time t_d at the point ρ , and that reaches us today at time $t_0 = 10^{10}$ years at $\rho = 0$. We then have

$$ds^2 = dt^2 - R(t)^2 \frac{d\rho^2}{1 - k\rho^2} = 0 \Rightarrow \frac{dt}{R(t)} = \pm \frac{d\rho}{\sqrt{1 - k\rho^2}}, \quad (12.1.17)$$

and thus

$$\int_{t_d}^{t_0} \frac{dt'}{R(t')} = \int_0^\rho \frac{d\rho'}{\sqrt{1 - k\rho'^2}}. \quad (12.1.18)$$

A second photon reaches us today from exactly the opposite direction. The coordinate difference between the two points of creation of the photons is therefore 2ρ . At the time t_d of their creation, this corresponded to a physical distance

$$\begin{aligned} d(t_d) &= 2 \int_0^\rho d\rho \sqrt{-g_{\rho\rho}} = 2 \int_0^\rho d\rho \frac{R(t_d)}{\sqrt{1 - k\rho^2}} = 2R(t_d) \int_{t_d}^{t_0} \frac{dt'}{R(t')} \\ &= 2d_H(t_d) \int_{t_d}^{t_0} \frac{dt'}{R(t')} / \int_0^{t_d} \frac{dt'}{R(t')}. \end{aligned} \quad (12.1.19)$$

Since the decoupling of the photons the Universe was matter dominated, i.e. $R(t) \propto t^{2/3}$. Thus we obtain

$$\begin{aligned} \frac{d(t_d)}{d_H(t_d)} &= 2(3t_0^{1/3} - 3t_d^{1/3})/3t_d^{1/3} \\ &= 2\left(\frac{t_0}{t_d}\right)^{1/3} - 2 \approx 2\left(\frac{10^{10}}{10^5}\right)^{1/3} \approx 40. \end{aligned} \quad (12.1.20)$$

This implies that the points at which the two photons were created were separated by 40 causality-lengths, and still the photons have the same temperature. Who has thermalized the two causally disconnected regions at extremely similar temperatures?

As we have seen earlier, a GUT phase transition unavoidably leads to the creation of magnetic monopoles via the Kibble mechanism. The contribution of monopoles to the energy density leads to $\Omega \approx 10^{11}$ in complete disagreement with observation. How can we get rid of the unwanted monopoles, without discarding GUTs altogether? After all, GUT were very useful, because they may explain the baryon asymmetry of the Universe?

12.2 The Idea of Inflation

The idea of the inflationary Universe is to use vacuum energy (a non-zero cosmological constant) to blow up the size of the Universe exponentially. When vacuum energy dominates the Friedmann equation takes the form

$$\frac{\dot{R}(t)^2}{R(t)^2} = \frac{\Lambda}{3} \Rightarrow R(t) = R(t_0) \exp\left(\sqrt{\frac{\Lambda}{3}}(t - t_0)\right). \quad (12.2.1)$$

When the inflationary period lasts for a time Δt the scale factor of the Universe increases by

$$Z = \frac{R(t_0 + \Delta t)}{R(t_0)} = \exp\left(\sqrt{\frac{\Lambda}{3}}\Delta t\right). \quad (12.2.2)$$

The inflationary period ends when the vacuum energy goes to zero, because some scalar field (e.g. of a GUT) rolls down to the stable minimum of a low energy broken phase. In that moment latent heat is released, and the Universe (now in the low energy broken phase) is reheated up to T_c . This process is accompanied by entropy and particle creation. The entropy increase is given by

$$\frac{S(t_0 + \Delta t)}{S(t_0)} = Z^3 = \exp(\sqrt{3\Lambda}\Delta t). \quad (12.2.3)$$

Assuming that the entropy before inflation was not unnaturally large (e.g. $S(t_0) \approx 1$), in order to explain $S \approx 10^{87}$ we need an inflation factor

$$Z = 10^{29} \Rightarrow \sqrt{\frac{\Lambda}{3}} \Delta t = 67. \quad (12.2.4)$$

An exponential expansion can also solve the horizon problem, because we then have

$$\begin{aligned} R(t_0 + \Delta t) \int_{t_0}^{t_0 + \Delta t} \frac{dt'}{R(t')} &= \frac{R(t_0 + \Delta t)}{R(t_0)} \frac{3}{\Lambda} [1 - \exp(-\sqrt{\frac{\Lambda}{3}} \Delta t)] \\ &\approx Z \sqrt{\frac{3}{\Lambda}} = \frac{Z}{\ln Z} \Delta t. \end{aligned} \quad (12.2.5)$$

For comparison, a radiation dominated Universe (with $R(t) \propto t^{1/2}$) has

$$R(\Delta t) \int_0^{\Delta t} \frac{dt'}{R(t')} = 2\Delta t. \quad (12.2.6)$$

Thus, the horizon problem is improved by a factor

$$\frac{Z}{2 \ln Z} = 10^{27}, \quad (12.2.7)$$

although a factor 40 would be sufficient to solve it.

Also the monopole problem is solved by inflation, because monopoles created before or during inflation are diluted by a factor Z^3 . In this way the monopole contribution to the critical density is reduced to

$$\Omega_M = 10^{11} Z^{-3} = 10^{-76}. \quad (12.2.8)$$

Hence, both the monopole and the horizon problem are solved, once the flatness (or entropy) problem is solved.

12.3 The Dynamics of Inflation

Let us consider a scalar field Φ (e.g. in a GUT) that obtains a vacuum expectation value via spontaneous symmetry breaking (e.g. to $SU(3) \otimes SU(2) \otimes U(1)$). The potential of the scalar field $V(\Phi)$ could have the standard Mexican hat shape. Let us assume that at time t_0 the system is in the symmetric phase with $\Phi = 0$.

Afterwards it slowly rolls down to the broken symmetry minimum with $|\Phi| = v$. The Lagrange density of the scalar field is given by

$$\mathcal{L}(\Phi, \partial_\mu \Phi) = \frac{1}{2} \partial^\mu \Phi \partial_\mu \Phi - V(\Phi), \quad (12.3.1)$$

and the corresponding energy momentum tensor is

$$T_{\mu\nu} = \partial_\mu \Phi \partial_\nu \Phi - \mathcal{L} g_{\mu\nu}. \quad (12.3.2)$$

Assuming that Φ is spatially constant, but time-dependent, for a Robertson-Walker metric one obtains

$$\begin{aligned} T_{00} &= \frac{1}{2} \dot{\Phi}^2 + V(\Phi) = \rho, \\ T_{ii} &= -(\frac{1}{2} \dot{\Phi}^2 - V(\Phi)) g_{ii} = -p g_{ii} \Rightarrow p = \frac{1}{2} \dot{\Phi}^2 - V(\Phi). \end{aligned} \quad (12.3.3)$$

As usual, we have identified density and pressure. The Einstein field equation

$$G_{\mu\nu} = 8\pi G T_{\mu\nu}, \quad (12.3.4)$$

then takes the following form

$$\begin{aligned} 3\left(\frac{\dot{R}^2}{R^2} + \frac{k}{R^2}\right) &= 8\pi G \left(\frac{1}{2} \dot{\Phi}^2 + V(\Phi)\right), \\ 2\frac{\ddot{R}}{R} + \frac{\dot{R}^2}{R^2} + \frac{k}{R^2} &= -8\pi G \left(\frac{1}{2} \dot{\Phi}^2 - V(\Phi)\right). \end{aligned} \quad (12.3.5)$$

We now identify the effective cosmological constant as

$$\Lambda = 8\pi G V(\Phi). \quad (12.3.6)$$

The typical scale for the potential $V(\Phi)$ is the GUT scale 10^{14} GeV. Hence, we can assume $V(\Phi) = 10^{14} \text{GeV}^4$ and thus

$$\sqrt{\frac{\Lambda}{3}} = \sqrt{\frac{8\pi}{3} \frac{V(0)}{m_P^2}} = \sqrt{\frac{8\pi}{3} \frac{10^{14}}{10^{19}}} 10^{14} \text{GeV} \approx 10^9 \text{GeV}. \quad (12.3.7)$$

In order to reach a sufficient amount of inflation, we need it to last for at least

$$\Delta t = 67 \sqrt{\frac{3}{\Lambda}} \approx 10^{-7} \text{GeV}^{-1} \approx 10^{-32} \text{sec}. \quad (12.3.8)$$

This is about 100-1000 times the age of the Universe at the time of inflation.

Is it possible to delay the transition to the broken phase with $\Phi = v$ for such a long time? To answer that question, we consider the equation of motion for the scalar field

$$\ddot{\Phi} + 3\frac{\dot{R}}{R}\dot{\Phi} + \frac{\partial V}{\partial \Phi} = 0. \quad (12.3.9)$$

The term $3(\dot{R}/R)\dot{\Phi}$, which deviates from the ordinary Klein-Gordon equation, results from the red-shift of the momentum of the scalar field. For a long-lasting inflationary period we need a small $\dot{\Phi}$. Ignoring $\ddot{\Phi}$ results in

$$\dot{\Phi} = -\frac{R}{3\dot{R}} \frac{\partial V}{\partial \Phi}. \quad (12.3.10)$$

In the Friedmann equation (with $k = 0$) we can furthermore neglect $\dot{\Phi}^2$ compared to $V(\Phi)$ such that

$$\frac{\dot{R}}{R} = \sqrt{\frac{8\pi G}{3}V(\Phi)} \Rightarrow \dot{\Phi} = -\frac{1}{\sqrt{24\pi G V(\Phi)}} \frac{\partial V}{\partial \Phi}. \quad (12.3.11)$$

The duration of inflation is now given by

$$\begin{aligned} \Delta t &= \int_{t_0}^{t_0+\Delta t} dt' = \int_{\Phi(t_0)}^{\Phi(t_0+\Delta t)} d\Phi \frac{1}{\dot{\Phi}} = \int_0^v d\Phi \sqrt{24\pi G V(\Phi)} \left(\frac{\partial V}{\partial \Phi}\right)^{-1} \\ &= \sqrt{24\pi} t_P \int_0^v d\Phi \sqrt{V(\Phi)} \left(\frac{\partial V}{\partial \Phi}\right)^{-1}. \end{aligned} \quad (12.3.12)$$

We parameterize the potential as

$$V(\Phi) = \lambda(\Phi^2 - v^2)^2, \quad (12.3.13)$$

and we obtain

$$\int_0^v d\Phi \sqrt{V(\Phi)} \left(\frac{\partial V}{\partial \Phi}\right)^{-1} = \int_0^v d\Phi \sqrt{\lambda} \frac{\Phi^2 - v^2}{2\lambda(\Phi^2 - v^2)2\Phi} = \frac{1}{4\sqrt{\lambda}} \int_0^v \frac{d\Phi}{\Phi} = \infty. \quad (12.3.14)$$

We need to distort Φ slightly from its unstable value $\Phi = 0$. For example, we can put it to $\Phi = \Phi_0$ initially, and obtain

$$\Delta t = \sqrt{\frac{3\pi}{2\lambda}} t_P \ln\left(\frac{v}{\Phi_0}\right). \quad (12.3.15)$$

The value of v is determined, for example, by the GUT scale, and $\lambda \approx 1$. In order to reach

$$\Delta t \approx 10^{-32} \text{sec} \approx 10^{12} t_P, \quad (12.3.16)$$

one needs an extremely small initial scalar field value

$$\Phi_0 \approx v \exp(-10^{12}). \quad (12.3.17)$$

In order to last long enough, the initial conditions for inflation must be adjusted extremely carefully. Hence, unfortunately, we have traded the fine-tuning problem for Ω for another one — the fine-tuning of Φ_0 . For other forms of the scalar potential the new fine-tuning problem may be less severe, but this scenario of inflation does not look very natural.

At the end of the inflationary period the scalar field approaches its vacuum value by oscillating around it. The oscillations are damped by the couplings to other fields. This leads to a release of the latent heat stored in the scalar field. The generated entropy reheats the Universe to the critical temperature T_c of the GUT transition.

12.4 Supernatural Inflation

The above model of inflation with a single scalar field rolling down slowly to its true minimum does not seem very natural, because it requires some very small parameters. In order to achieve a sufficient number of e-foldings, one needs a potential $V(\Phi)$ which is very flat around $\Phi = 0$. Flat directions are unnatural and require fine-tuning of parameters in ordinary quantum field theories. However, in supersymmetric theories flat directions may arise naturally. Still, for a successful inflationary model some small parameters are required. In the framework of supernatural inflation one avoids to introduce small parameters beyond the ones that already exist in particle physics. For example, although it is not understood why the up-quark Yukawa coupling is so small, one can use this as a small number that certainly exists in the theory before one discusses inflation. Also the ratio of the electroweak and the GUT scale is a candidate for a small parameter, which exists in particle physics, even though we don't understand its origin. Supernatural inflation uses these kind of small parameters in the scalar potential. In addition, one works with two scalar fields, one that provides the vacuum energy for inflation, and another that controls the ending of inflation.

First, let us again look at models with one scalar field Φ and with a potential that we parameterize as

$$V(\Phi) = M^4 g(\Phi/v). \quad (12.4.1)$$

Here M is a typical symmetry breaking scale, for example the GUT scale, v is the vacuum value of Φ , and $g(x)$ is some dimensionless function of order 1. The

time that inflation lasts in such a potential is again given by

$$\Delta t = \int_{t_0}^{t_0+\Delta t} dt' = -\sqrt{24\pi}t_P \int_{\Phi_0}^v d\Phi \sqrt{V(\Phi)} \left(\frac{\partial V}{\partial \Phi}\right)^{-1}. \quad (12.4.2)$$

During the inflationary epoch we have an effective cosmological constant $\Lambda = 8\pi G V(\Phi)$ which is essentially constant. Hence, the number of e-foldings results in

$$\sqrt{\frac{\Lambda}{3}} \Delta t = -\frac{8\pi}{M_P^2} \int_{\Phi_0}^v d\Phi V(\Phi) \left(\frac{\partial V}{\partial \Phi}\right)^{-1} \approx -\frac{8\pi}{M_P^2} (v - \Phi_0) v g \left(\frac{\partial g}{\partial x}\right)^{-1}. \quad (12.4.3)$$

Since g is of order one, one needs v to be of the order of the Planck mass M_P . This is not a scale that is natural in a particle physics context. In particular, the typical scales of supersymmetric theories, which naturally have flat directions, are much below M_P . Also, inflation before a GUT phase transition would not solve the monopole problem.

Supernatural inflation works with two scalar fields Φ and Ψ . A simple form of the scalar potential with the desired features is

$$V(\Phi, \Psi) = \lambda(\Phi^2 - v^2)^2 + \frac{m^2}{2}\Psi^2 + \lambda'\Phi^2\Psi^2. \quad (12.4.4)$$

The mass m of the scalar field Ψ is very small (in the TeV range), which is natural in supersymmetric theories. Hence, for small values of Φ the potential is very flat in the Ψ -direction. In this model, at the beginning of inflation Ψ has a large value, and Φ is close to zero. Then Ψ is rolling very slowly towards smaller values. Slow rolling is natural without fine-tuning, because a small m is natural due to supersymmetry. As long as Ψ is large, i.e. as long as

$$\lambda'\Psi^2 > 2\lambda v^2, \quad (12.4.5)$$

the scalar field Φ has a positive mass squared, and remains at $\Phi = 0$. Then λv^4 provides the vacuum energy for inflation. Once Ψ becomes too small, Φ gets a negative mass squared. Then it rolls down to its absolute minimum at $\Phi = v$ and inflation comes to an end. The duration of inflation in this model is given by

$$\Delta t = \int_{t_0}^{t_0+\Delta t} dt' = \int_{\Psi_0}^{v\sqrt{2\lambda/\lambda'}} d\Psi \frac{1}{\Psi}. \quad (12.4.6)$$

The equation of motion for the slowly rolling Ψ field (with $\Phi = 0$) is

$$\dot{\Psi} = -\frac{R}{3R} \frac{\partial V}{\partial \Psi} = -\frac{1}{\sqrt{3}\Lambda} m^2 \Psi. \quad (12.4.7)$$

The effective cosmological constant is given by $\Lambda = 8\pi G\lambda v^4$, such that the number of e-foldings now is

$$\sqrt{\frac{\Lambda}{3}}\Delta t = -\frac{8\pi}{M_P^2} \int_{\Psi_0}^{v\sqrt{2\lambda/\lambda'}} d\Psi \frac{1}{m^2\Psi} = \frac{8\pi\lambda v^4}{M_P^2 m^2} \ln \frac{\Pi_0}{v\sqrt{2\lambda/\lambda'}}. \quad (12.4.8)$$

Hence, for supernatural inflation v need not be at the Planck scale. Instead $v \approx \sqrt{M_P m}$, where m is very small compared to M_P . Still, in order to generate a reasonable spectrum of density fluctuations in this model one must assume that $\lambda' \approx 10^{-8}$. This seems again very unnatural. However, one can argue that λ' is related to a small number that already exists in the standard model — namely the Yukawa coupling of the up-quark. When one gives up renormalizability of the scalar potential, one can even completely avoid small dimensionless numbers. For example, in the potential

$$V(\Phi, \Psi) = \lambda(\Phi^2 - v^2)^2 + \frac{m^2}{2}\Psi^2 + \frac{1}{M'^2}(\Phi^4\Psi^2 + \Phi^2\Psi^4) \quad (12.4.9)$$

the parameter λ' is replaced by a mass scale M' , which could for example be the GUT scale. Since, fine-tuning is eliminated, supernatural inflation models appear much more natural, thanks to the flat directions in supersymmetric theories.

In conclusion, the idea of inflation is very attractive, because it cures several deficiencies of the standard cosmology, and explains some of its initial conditions. It remains to be seen if the corresponding dynamics of the scalar field can be embedded in a realistic particle physics model. A main prediction of inflation is $\Omega = 1$. Together with the bound on the baryonic energy density $\Omega_{baryons} \leq 0.18$ resulting from primordial nucleosynthesis calculations, this implies that 80 percent of the existing matter is of non-baryonic origin. If the right amount of dark matter would be found, this would be a great triumph for the idea of inflation. Inflation also predicts a spectrum of density fluctuations that may give rise to the small anisotropies in the cosmic background radiation, and that may explain the large scale structures that we observe in the Universe today. Again, if models of inflation will be successful in generating the right spectrum of initial fluctuations, the idea of inflation could be promoted to an established fact.

Chapter 13

Quantum Cosmology

Up to this point we have treated gravity classically. This was justified, because we have limited ourselves to times later than the Planck time $t_P \approx 10^{-44}$ sec. At present, we cannot talk about earlier times with some confidence, because we have no well-established theory of quantum gravity. Still, many interesting ideas are around, and nobody can stop us to speculate about the very earliest moments in the history of our Universe. The idea of inflation applies to times after the Planck time, and makes the later Universe insensitive to the earliest initial conditions. Hence, it may be impossible to extrapolate back to t_P from the present epoch. In any case, it is unlikely that any idea about quantum cosmology will soon be tested in observations.

To apply quantum mechanics to the Universe as a whole is, on the one hand, paradoxical, on the other hand it seems necessary. It is necessary, if we postulate that quantum mechanics is universally valid on all scales — not just microscopically. In addition, at the Planck time the observable part of the Universe was itself microscopically small. In this chapter we will discuss the Wheeler-DeWitt equation, whose solution is the “wave function of the Universe”. On the other hand, the concept of a wave function for the whole Universe is paradoxical. After all, a wave function contains information about the statistical distribution of results of measurements. If the object of study is the whole Universe, who is going to do the measurements? Also, to test the probabilistic predictions of quantum mechanics one always needs to do repeated measurements on several identically prepared physical systems. The wave function of the Universe could be compared with “experiments” only if we could perform measurements on several identically prepared Universes. An observer inside the Universe, who is part of the quantum system that he wants to study, will not be able to probe the wave function of the

Universe.

This leads us to another problem that arises when we apply quantum mechanics to the Universe as a whole. Up to now we have discussed various aspects of the history of the Universe. Classically, this makes a lot of sense. Quantum mechanically, however, history arises only from measurements. As long as no measurements are performed, the wave function just follows the Schrödinger equation. This means that different events can happen with different probabilities. Once a measurement is performed, the system “makes up its mind”, and “chooses” a given event from the appropriate probability distribution — thus turning it into history. When quantum mechanics is applied to the whole Universe, who has made the “measurements” that turned the cosmological events we have discussed earlier into the history of our Universe? These questions show that, in any case, quantum mechanics is not understood well enough, to apply it to a system that includes the observer. For us, the Universe is certainly a system of that kind. Whatever comes out of quantum cosmology calculations, we will have a hard time interpreting the results.

General relativity can be formulated in the Hamiltonian formulation. The generalized “coordinates” then are the possible spatial metrics and the matter field configurations. The wave function of the Universe is a functional of these coordinates, which span so-called “superspace”, which in this case is totally unrelated to supersymmetry. One can construct the canonical conjugate momenta for the generalized coordinates, and then use canonical quantization of the system. This leads to a non-renormalizable quantum theory of infinitely many non-linearly coupled degrees of freedom. A tractable problem can be formulated in so-called “mini-superspace”. One then limits oneself to a single degree of freedom, e.g. the scale parameter of the Universe. When combined with the constant mode of the scalar field driving inflation, one can perhaps learn something about the initial conditions for inflation from this model.

The most promising ideas about quantum gravity are based on string theory, which is defined in higher dimensions. Like in Kaluza-Klein theories, all but the four physical dimensions are compactified, and have a small extent of the order of the Planck length. In the ultra-early Universe, however, also the extent of the other four directions was tiny, and there may have been an interplay between the compactified dimensions and space-time. Needless to say that these ideas are again highly speculative. Still, as long as we take this not too seriously, it is interesting to pursue such a scenario.

13.1 Wheeler-DeWitt Equation in Mini-Superspace

Let us consider a positively curved Universe with scale parameter $R(t)$. The only source of the energy density is assumed to be a cosmological constant Λ . Then, the Friedmann equation takes the form

$$\left(\frac{\dot{R}}{R}\right)^2 + \frac{1}{R^2} = \frac{\Lambda}{3}. \quad (13.1.1)$$

This equation is solved by

$$R(t) = \sqrt{\frac{\Lambda}{3}} \cosh\left(\sqrt{\frac{3}{\Lambda}}t\right), \quad (13.1.2)$$

i.e. the Universe contracts to a minimal radius and then expands forever.

In this case, the Einstein-Hilbert action takes the form

$$\begin{aligned} S &= -\frac{1}{16\pi G} \int d^4 \sqrt{-g} (\mathcal{R} + 2\Lambda) \\ &= -\frac{3\pi}{4G} \int dt (R\dot{R}^2 - R + R^3 \frac{\Lambda}{3}). \end{aligned} \quad (13.1.3)$$

Hence, we can read off the classical Lagrange function

$$L(R, \dot{R}) = -\frac{3\pi}{4G} (R\dot{R}^2 - R + R^3 \frac{\Lambda}{3}). \quad (13.1.4)$$

The canonical conjugate momentum to the variable R is

$$P = \frac{\partial L}{\partial \dot{R}} = -\frac{3\pi}{2G} R\dot{R}, \quad (13.1.5)$$

and the classical Hamilton function is given by

$$\begin{aligned} H &= P\dot{R} - L \\ &= -\frac{3\pi}{4G} (R\dot{R}^2 + R - R^3 \frac{\Lambda}{3}) \\ &= -\frac{G}{3\pi R} P^2 - \frac{3\pi}{4G} (R - R^3 \frac{\Lambda}{3}). \end{aligned} \quad (13.1.6)$$

Inserting the classical solution from above, one obtains $H = 0$. This is not just a special case for this particular solution. Instead, $H = 0$ arises as a constraint from the gauge structure of general coordinate transformations underlying general relativity.

One now follows a canonical quantization prescription and replaces

$$P = i \frac{\partial}{\partial R}, \quad (13.1.7)$$

such that the above Hamilton function turns into the Hamilton operator

$$H = \frac{G}{3\pi R} \frac{\partial^2}{\partial R^2} - \frac{3\pi}{4G} (R - R^3 \frac{\Lambda}{3}). \quad (13.1.8)$$

The classical constraint $H = 0$ cannot be satisfied as an operator identity. Instead, one implements it as a constraint on the states of the theory, i.e.

$$H\Psi(R) = 0. \quad (13.1.9)$$

This is analogous to the Gauss law constraint in an ordinary gauge theory. Inserting the specific form of the Hamiltonian one obtains

$$[\frac{\partial^2}{\partial R^2} - \frac{9\pi^2}{4G^2} (R^2 - R^4 \frac{\Lambda}{3})] \Psi(R) = 0. \quad (13.1.10)$$

This is the Wheeler-DeWitt equation in mini-superspace. One can view it as a Schrödinger equation for zero-energy wave functions. Interestingly, the “wave function of the Universe” $\Psi(R)$ is independent of time. Since the interpretation of this wave function is unclear, we also do not understand what it means that it is time-independent.

13.2 The Wave Function of the Universe

The above Wheeler-DeWitt equation corresponds to a zero-energy Schrödinger equation in the potential

$$U(R) = \frac{9\pi^2}{4G^2} (R^2 - R^4 \frac{\Lambda}{3}). \quad (13.2.1)$$

The potential is positive in the classically forbidden region $R < \sqrt{\Lambda/3}$. Indeed, classically the considered Universe has a minimal radius $\sqrt{\Lambda/3}$. Quantum mechanically, the Universe can now tunnel to the classically forbidden regime. Indeed, at $R = 0$ the potential energy is again zero. One can also start from this point and tunnel to the classical region. The point $R = 0$ corresponds to a Universe of zero size — really just “nothing”. Quantum mechanically, we can thus imagine to make a classical expanding Universe out of “nothing” via a tunneling transition.

Since the physical interpretation of the wave function of the Universe is unclear, one also has problems selecting appropriate boundary conditions. Without a choice of boundary conditions the Wheeler-DeWitt equation has many solutions, and it is an open question which one (if any) is physical.

Let us follow Hartle's and Hawking's choice of boundary conditions and compute the tunneling rate for creating a classical Universe from nothing. The tunneling rate is given by the Boltzmann factor for the Euclidean action for a classical tunneling path $\exp(-S_E)$. The Euclidean action is given by

$$S_E = -\frac{3\pi}{4G} \int dt (R\dot{R}^2 + R - R^3 \frac{\Lambda}{3}), \quad (13.2.2)$$

and the corresponding Euclidean equation of motion is

$$(\frac{\dot{R}}{R})^2 - \frac{1}{R^2} = -\frac{\Lambda}{3}. \quad (13.2.3)$$

The classical tunneling solution is

$$R(t) = \sqrt{\frac{\Lambda}{3}} \cos(\sqrt{\frac{3}{\Lambda}} t), \quad (13.2.4)$$

and the Euclidean action takes the form

$$S_E = -\frac{3\pi}{2G} \int dt R\dot{R}^2 = -\frac{3\pi}{G\Lambda}. \quad (13.2.5)$$

Hence the tunneling probability for the creation of a classical Universe is

$$P_t \propto \exp(-S_E) = \exp(\frac{3\pi}{G\Lambda}). \quad (13.2.6)$$

The tunneling rate is larger for a large vacuum energy. It should, however, be mentioned that other choices of boundary conditions lead to different results.

13.3 The Initial Conditions for Inflation

Let us now discuss an extension of the mini-superspace model. Instead of assuming a cosmological constant, we now assume a spatially constant scalar field Φ , whose potential $V(\Phi)$ provides vacuum energy when Φ is displaced from its vacuum expectation value. The field Φ could be the one that drives inflation.

Thus, we are interested in the initial conditions for inflation provided by quantum cosmology at the Planck time. The Einstein-Hilbert action with the scalar field takes the form

$$\begin{aligned} S &= \int d^4 \sqrt{-g} \left(-\frac{1}{16\pi G} \mathcal{R} + \frac{1}{2} \partial^\mu \Phi \partial_\mu \Phi - V(\Phi) \right) \\ &= 2\pi^2 \int dt \left(-\frac{6R\dot{R}^2 - 6R}{16\pi G} + R^3 \left(\frac{1}{2} \dot{\Phi}^2 - V(\Phi) \right) \right). \end{aligned} \quad (13.3.1)$$

The conjugate momentum of the scalar field is

$$\Pi = 2\pi^2 R^3 \dot{\Phi}, \quad (13.3.2)$$

and the extended Hamilton function takes the form

$$\begin{aligned} H &= P\dot{R} + \Pi\dot{\Phi} - L \\ &= -\frac{G}{3\pi R} P^2 + \frac{1}{4\pi^2 R^3} \Pi^2 - \frac{3\pi}{4G} R + 2\pi^2 R^3 V(\Phi). \end{aligned} \quad (13.3.3)$$

Again, we proceed via canonical quantization and replace

$$P = i \frac{\partial}{\partial R}, \quad \Pi = i \frac{\partial}{\partial \Phi}. \quad (13.3.4)$$

The Wheeler-DeWitt equation then takes the form

$$\left[\frac{\partial^2}{\partial R^2} - \frac{3}{4\pi G R^2} \frac{\partial^2}{\partial \Phi^2} - U(R, \Phi) \right] \Psi(R, \Phi) = 0. \quad (13.3.5)$$

The potential entering the effective zero-energy Schrödinger equation is now given by

$$U(R, \Phi) = \frac{9\pi^2}{4G^2} \left(R^2 - R^4 \frac{8\pi G V(\Phi)}{3} \right). \quad (13.3.6)$$

As expected, $8\pi G V(\Phi)/3$ plays the role of the cosmological constant. Repeating the tunneling calculation following Hartle and Hawking, one finds that the amplitude is largest for tunneling to a configuration of Φ that has the largest $V(\Phi)$, i.e. the largest cosmological constant. This is indeed welcome, because it may explain how the scalar field started away from its vacuum value before inflation. However, again it should be noted that these results change when one uses different boundary conditions for the wave function of the Universe.

13.4 Cosmology with Extra Dimensions

Motivated by string and Kaluza-Klein theories, some people are studying cosmology in higher dimensional Universes with compactified directions. For example,

let us consider a theory in D dimensions, with a flat four-dimensional space-time with a scale parameter $R(t)$ and $D-4$ extra dimensions which form a torus of side length $S(t)$. The Einstein tensor for such a metric has the following components

$$\begin{aligned} G_{00} &= -3\frac{\ddot{R}}{R} - (D-4)\frac{\ddot{S}}{S}, \\ G_{ij} &= -\left[\frac{\ddot{R}}{R} + 2\frac{\dot{R}^2}{R^2} + (D-4)\frac{\dot{R}}{R}\frac{\dot{S}}{S}\right]\delta_{ij}, \\ G_{\alpha\beta} &= -\left[\frac{\ddot{S}}{S} + (D-5)\frac{\dot{S}^2}{S^2} + 3\frac{\dot{R}}{R}\frac{\dot{S}}{S}\right]\delta_{\alpha\beta}. \end{aligned} \quad (13.4.1)$$

Here i, j are spatial indices, while α, β denote indices of the extra dimensions.

Of course, one can consider such a Universe with any kind of matter content. For simplicity, we consider an empty Universe, i.e. one with vanishing energy-momentum tensor. Then the Einstein equations of motion take the form

$$\begin{aligned} 3\frac{\ddot{R}}{R} + (D-4)\frac{\ddot{S}}{S} &= 0, \\ \frac{\ddot{R}}{R} + 2\frac{\dot{R}^2}{R^2} + (D-4)\frac{\dot{R}}{R}\frac{\dot{S}}{S} &= 0, \\ \frac{\ddot{S}}{S} + (D-5)\frac{\dot{S}^2}{S^2} + 3\frac{\dot{R}}{R}\frac{\dot{S}}{S} &= 0. \end{aligned} \quad (13.4.2)$$

Let us make a power-law ansatz for the two scale factors

$$R(t) = R(t_0)\left(\frac{t}{t_0}\right)^n, \quad S(t) = S(t_0)\left(\frac{t}{t_0}\right)^m. \quad (13.4.3)$$

The above equations then take the form

$$\begin{aligned} 3n(n-1) + (D-4)m(m-1) &= 0, \\ n(n-1) + 2n^2 + (D-4)nm &= 0, \\ m(m-1) + (D-5)m^2 + 3nm &= 0. \end{aligned} \quad (13.4.4)$$

They are solved by

$$n = \frac{1 + \sqrt{(D-2)(D-4)}/3}{D-1}, \quad m = \frac{D-4 - \sqrt{3(D-2)(D-4)}}{(D-1)(D-4)}. \quad (13.4.5)$$

Without the help of the extra dimensions, an empty Universe would follow

$$3\frac{\ddot{R}}{R} = 0, \quad \frac{\ddot{R}}{R} + 2\frac{\dot{R}^2}{R^2} = 0, \quad (13.4.6)$$

which leads to $R(t) \propto t$ and thus $n = 1$. With the extra dimensions involved, the Universe expands at a different rate, depending on the number of extra dimensions. The additional dimensions shrink and finally disappear, because

$$D - 4 < \sqrt{3(D - 2)(D - 4)}. \quad (13.4.7)$$

In an expanding higher-dimensional Universe the entropy in a co-moving higher dimensional volume is conserved. The volume of the compactified directions, however, is rapidly decreasing, and thus the effective entropy of the physical three-dimensional Universe seems to increase. This is interesting, because inflation solved the flatness or age problem by providing the enormous entropy 10^{87} . Detailed studies of cosmology in higher dimensions indicate that the mechanism of providing entropy from higher dimensions is not sufficient to provide such a huge factor. Still, it is interesting that cosmological solutions with three large expanding and $D - 4$ small shrinking dimensions at least exist.

Chapter 14

Large Scale Structure

Although the Universe seems to be homogeneous and isotropic on the largest scales, on smaller (but still very large) scales a lot of structures have emerged during the evolution of the Universe. There are galaxies, galaxy clusters, superclusters, as well as large voids. The isotropy of the cosmic background radiation, which reflects the situation at about 10^5 years after the bang, indicates that the initial fluctuations must have been very small. Can we understand how the structure that we see today have emerged from these small initial fluctuations? We will consider this question at a qualitative level only. To answer it quantitatively requires large scale numerical work. Structure formation is a question of the correct initial conditions: what is the origin of the initial fluctuations? One distinguishes two cases: quantum fluctuations of the scalar field that drives inflation, and density fluctuations due to a network of cosmic strings. Structure formation is also significantly influenced by the type of dark matter one assumes. In any case, there is at least 10 times more dark than shining matter. Again, one can distinguish two cases: hot and cold dark matter. Hot dark matter consists of relativistic particles like massive (but very light) neutrinos or axions. Cold dark matter is non-relativistic. Candidates for cold dark matter are WIMPs (weakly interacting massive particles), e.g. supersymmetric particles like the gravitino or the photino, or MACHOs (massive compact halo objects), but also black holes. Numerical simulations of hot dark matter seem to yield realistic structures, but structure formation sets in rather late, such that the existence of very old galaxies cannot be explained. Cold dark matter has its own problems: the corresponding numerical simulations do not yield the large voids that one observes. Also the best results are obtained with $\Omega \approx 0.2$, not with $\Omega = 1$ as suggested by inflation.

Up to this point, we have reconstructed the history of the Universe. Based on

what we know about particle physics, we were able to talk about times as early as 10^{-12} sec after the big bang with some confidence. Using more speculative ideas (GUT), we could extend our considerations to times down to 10^{-34} sec, and we speculated rather wildly even about the physics at the Planck time, 10^{-44} sec. In the last section of this last chapter, we will use the same ideas as before to predict the future of the Universe. We will talk about times between today (10^{10} years after the bang) and the very distant future 10^{140} years after the bang. By then the Universe has either disappeared, or has become a rather dull place. Or maybe it has changed completely to another vacuum with new interesting physics. In any case, we won't know for sure within our life-time and neither will any other human being born some time in the future. Since our ideas about the far future are not testable in experiments or observations, one may argue that they are not part of physics. Still, it is interesting to speculate what will happen to the Universe when we have long disappeared.

14.1 The Jeans Instability

What is the typical scale at which we should expect structure formation in a gravitating system? To answer this question we consider a non-relativistic gas of matter, which for simplicity we treat using Newton's theory of gravity. Of course, in an expanding Universe the results will be modified by general relativity. An ideal gas of matter (without friction, viscosity) obeys the following equations

$$\begin{aligned}\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot (\rho \vec{v}) &= 0, \\ \frac{\partial \vec{v}}{\partial t} + (\vec{v} \cdot \vec{\nabla}) \vec{v} + \frac{1}{\rho} \vec{\nabla} p + \vec{\nabla} \Phi &= 0, \\ \Delta \Phi &= 4\pi G \rho.\end{aligned}\tag{14.1.1}$$

Here ρ and p are density and pressure of the gas, \vec{v} is its velocity field, and Φ is Newton's gravitational potential. The first equation is the continuity equation, the second one is the Navier-Stokes equation, and the third equation determines the gravitational field. Now we consider small deviations $\tilde{\rho}, \tilde{p}, \tilde{\vec{v}}, \tilde{\Phi}$ around a static and spatially constant situation with values $\bar{\rho}, \bar{p}, \bar{\vec{v}}, \bar{\Phi}$, i.e.

$$\rho = \bar{\rho} + \tilde{\rho}, \quad p = \bar{p} + \tilde{p}, \quad \vec{v} = \bar{\vec{v}} + \tilde{\vec{v}}, \quad \Phi = \bar{\Phi} + \tilde{\Phi}.\tag{14.1.2}$$

Linearizing the equations of motion in the small deviations, one obtains

$$\begin{aligned}\frac{\partial \tilde{\rho}}{\partial t} + \bar{\rho} \vec{\nabla} \cdot \tilde{\vec{v}} &= 0, \\ \frac{\partial \tilde{\vec{v}}}{\partial t} + \frac{1}{\bar{\rho}} \vec{\nabla} \tilde{p} + \vec{\nabla} \tilde{\Phi} &= 0, \\ \Delta \tilde{\Phi} &= 4\pi G \tilde{\rho}.\end{aligned}\tag{14.1.3}$$

We have five equations for six unknowns. The missing relation is the equation of state

$$\tilde{p} = \tilde{\rho} v_s^2.\tag{14.1.4}$$

Here v_s is the velocity of sound. We then get

$$\frac{\partial \tilde{\vec{v}}}{\partial t} + \frac{v_s^2}{\bar{\rho}} \vec{\nabla} \tilde{\rho} + \vec{\nabla} \tilde{\Phi} = 0.\tag{14.1.5}$$

Taking another time-derivative of the continuity equation we find

$$\begin{aligned}\frac{\partial^2 \tilde{\rho}}{\partial t^2} + \bar{\rho} \vec{\nabla} \cdot \frac{\partial \tilde{\vec{v}}}{\partial t} &= \frac{\partial^2 \tilde{\rho}}{\partial t^2} + \bar{\rho} \vec{\nabla} \cdot \left(-\frac{v_s^2}{\bar{\rho}} \vec{\nabla} \tilde{\rho} - \vec{\nabla} \tilde{\Phi} \right) \\ &= \frac{\partial^2 \tilde{\rho}}{\partial t^2} - v_s^2 \Delta \tilde{\rho} - \bar{\rho} \Delta \tilde{\Phi} \\ &= \frac{\partial^2 \tilde{\rho}}{\partial t^2} - v_s^2 \Delta \tilde{\rho} - 4\pi G \bar{\rho} \tilde{\rho} = 0.\end{aligned}\tag{14.1.6}$$

Performing a Fourier transform

$$\tilde{\rho}(\vec{k}, \omega) = \int d^3x \int dt \tilde{\rho}(\vec{x}, t) \exp(i(\vec{k} \cdot \vec{x} - \omega t)),\tag{14.1.7}$$

one obtains the dispersion relation

$$-\omega^2 + v_s^2 k^2 - 4\pi G \bar{\rho} = 0 \Rightarrow \omega = \pm \sqrt{v_s^2 k^2 - 4\pi G \bar{\rho}}.\tag{14.1.8}$$

Real values of ω correspond to density oscillations (sound waves), while an imaginary ω corresponds to exponentially growing or decreasing perturbations. The critical wave number that separates the two regimes is the Jeans wave number

$$k_J = \sqrt{\frac{4\pi G \bar{\rho}}{v_s^2}}.\tag{14.1.9}$$

Spatial fluctuations of large wave number ($k > k_J$) have a real ω and correspond to oscillations. Long wave length fluctuations with $k < k_J$ have purely imaginary

ω and can thus grow exponentially. Let us also define the Jeans mass as the total mass within a sphere of radius π/k_J

$$M_J = \frac{4\pi}{3} \left(\frac{\pi}{k_J} \right)^3 \bar{\rho} = \frac{\pi^{5/2}}{6} \frac{v_s^3}{G^{3/2} \bar{\rho}^{1/2}}. \quad (14.1.10)$$

Fluctuations of smaller mass are oscillatory, and hence stable against gravitational collapse, while masses $M > M_J$ are unstable and may thus form large scale structures.

When the above analysis is extended to an expanding Universe using general relativity, one obtains a Jeans mass of about 10^6 solar masses. Galaxies, on the other hand, often have 10^{11} solar masses. The Jeans scale therefore does not directly explain the size of galaxies, it only provides a lower bound for structure formation.

14.2 Four Scenarios for Structure Formation

At present there is no standard model for structure formation. Instead there are various scenarios, which essentially decompose into four classes. They are distinguished by different assumptions about the origin of the initial density fluctuations, as well as about the form of dark matter. The initial fluctuations are assumed to be either due to the quantum fluctuations of the scalar field that drives inflation, or due to a network of cosmic strings. Also one distinguishes between hot and cold dark matter. Hot dark matter is still relativistic today, and could, for example, consist of light (but not massless) neutrinos. Cold dark matter, on the other hand, is nonrelativistic and could consist of WIMPs or MA-CHOs. Combining the various options we obtain four basic scenarios for structure formation. Models with initial fluctuations coming from inflation are presently better understood, and seem to work better than those based on cosmic strings.

In case of hot dark matter (neutrinos) with fluctuations due to inflation, density fluctuations of the neutrinos lead to large structures (super-clusters) which collapse to discs (pancakes), and then differentiate further to smaller structures like galaxies. This, however, sets in so late, that in this scenario the origin of old galaxies remains unclear.

With cold dark matter small structures like galaxies form first, and hence one has no problems with old galaxies. However, in simulations of cold dark matter one does not find the large voids, that one observes in the Universe. In addition, the simulations favor $\Omega \approx 0.2$ not $\Omega = 1$ as predicted by inflation.

Cosmic strings can exist only if they are generated after inflation. The initial structures simply form along the cosmic string network, which acts as a seed for galaxy formation. Numerical simulations of cosmic strings and cold dark matter seem to generate large structures successfully. A fluctuating cosmic string network constantly emits gravitational waves. The resulting background of gravitational waves should be detectable in the very precise measurements of the double quasar PSR 1937+21. The fourth scenario — cosmic strings with hot dark matter — is presently not very well understood.

In conclusion, there are some promising scenarios for structure formation which can explain the observed structures more or less successfully. The problem is to generate large structures from the small initial fluctuations that are reflected by the highly isotropic cosmic background radiation. Did the Universe have enough time during 10^{10} years of expansion to turn from the hot, uniform, structureless gas of elementary particles in the early Universe into the highly structured system in which we live today?

14.3 Four Eras in the Late Universe

It is interesting to ask what will happen to the Universe in the distant future, even if we won't be able to test these ideas in observations. Whatever we predict will be based on our incomplete knowledge about particle physics (and physics in general) today. Thus, many of the following considerations may have to be revised, as our knowledge progresses.

Following Adams and Laughlin, we divide the future somewhat arbitrarily into four eras. During the stelliferous era that started with galaxy formation, stars dominate the Universe. Stars burn out after a while, and star formation will come to an end at about 10^{14} years after the bang. By then, the stellar objects have turned into brown dwarfs, white dwarfs, neutron stars, and an occasional black hole. White dwarfs and neutron stars are supported by degeneracy pressure of electrons and neutrons, respectively. When these objects dominate the Universe, we enter the degenerate era, which lasts until about 10^{40} years after the bang. At that time, all protons will have decayed, and thus all stellar objects have disappeared. As a result, we enter the black hole era, which ends when the last biggest black hole in the Universe has evaporated at about 10^{100} years. After that, in the dark era, the Universe consists of photons, as well as some electrons and positrons, which ultimately annihilate into photons at about 10^{140} years after the bang. If nothing else happens by then, the Universe will be a very dull place,

unpleasant to live in for even the most patient creatures.

Let us start discussing the stelliferous era, which we have just entered. The crucial observation is that, while massive stars dominate the Universe now, they will soon run out of fuel. Depending on their mass, they will either blow up in a supernova, leaving behind a neutron star, or they will become a red giant, and then end as a white dwarf. Less massive stars, on the other hand, burn the nuclear fuel at a much more moderate rate. These objects therefore live much longer, up to 10^{13} years, and finally also become white dwarfs (without going through a red giant phase). There are also brown dwarfs, objects which are not massive enough to ignite a nuclear fire in their cores. They remain unchanged during the stelliferous era.

The fate of our earth is decided when our sun enters the red giant phase in about five billion years. Most likely the earth will be swallowed and grilled within a time of about 50 years. Still, there is a chance that the sun loses enough material before going into the red giant phase, thus shifting the planets to further away orbits. This may save the earth for a long time to come.

Star formation happens in the gaseous clouds within galaxies. However, the supply of material is finite, and after a while there will be nothing left to make stars. One can estimate that the last star will form at about 10^{14} years after the bang, and it will soon be burned out. This is when the stelliferous era comes to an end.

At the end of the stelliferous era we are left with a system of many brown and white dwarfs, and an occasional neutron star or black hole, and we now enter the degenerate era. The above stellar objects are still bound inside galaxies, but galaxies will now die, because over time the stellar objects evaporate at a time scale of 10^{20} years, or fall into the black hole at the center of the galaxy. Still, during this era, occasionally a luminous star may be formed in a collision of brown or white dwarfs. One can estimate the collision rate for such processes, and one finds that a typical galaxy will then contain about 50 luminous stars, compared to 10^{10} today. At about 10^{30} years after the bang, all galaxies will have turned into a system of black holes and single evaporated stars.

The degenerate era comes to an end when the evaporated stellar objects outside black holes disappear due to proton decay. Depending on one's favorite GUT theory, this may happen at different times, e.g. 10^{37} years after the bang. When a proton decays, it can make a pion and a positron. The positron annihilates with an electron and makes two photons, and the pion also decays into two photons. At the end we lose a proton and an electron, and we get four photons.

This means that the stars continue to shine, however, with a luminosity of about 10^{-24} of the sun. The stellar temperature then is about 0.06 K. One such star can power no more than a couple of light bulbs. Before a white dwarf disappears completely, its electron degeneracy is lifted, because the mass gets too small. Similarly, a neutron star will turn into a white dwarf, when sufficiently many neutrons have decayed. Also planets — e.g. the earth if it survived the sun's red giant phase — will disappear via proton decay, thus emitting a few photons. The luminosity, however, is ridiculously small, about 0.4 mW. The degenerate era may come to an end at about 10^{40} years after the bang, assuming the above life-time for the nucleon. Based on the standard model alone — i.e. relying on sphaleron processes — the proton would live much longer, about 10^{172} years.

At the end of the degenerate era we enter the black hole era, during which black holes coalesce and finally disappear by emitting Hawking radiation at about 10^{100} years after the big bang. We then enter the dark era, in which the universe consists of just photons, and a few electrons and positrons. Electrons and positrons form positronium states — typically of the horizon scale, and cascade down to the ground state, from which they soon annihilate and become photons. Then, at about 10^{140} years after the big bang, not much can happen, because there is nobody left who would care about anybody else. Well, perhaps the photons like to go into a Bose condensate and enjoy themselves for the rest of time.

Of course, for the Universe to survive until 10^{140} years after the bang, we have assumed that the Universe is not closed. Otherwise, the fun might end as early as 10^{11} years. Even if the Universe is flat in the part that we can observe now, as it expands a large density fluctuation may enter the horizon later, and close it. Of course, we can again use inflation to make the Universe homogeneous on scales even larger than the ones we care about today. In order to guarantee survival of the Universe for about 10^{100} years, we need about 130 e-foldings, which is perhaps not impossible with sufficiently clever inflationary model building.

Also, any negligibly small amount of vacuum energy today, will dominate the Universe in the distant future. This could lead to another inflationary epoch. Thus, in order to go through the previously discussed eras, we must assume an even tinier cosmological constant. Otherwise, stellar objects may lose causal contact before they disappear via proton decay. We would see stars disappear through the horizon, and it gets dark much before the dark era.

An even more drastic effect may be due to vacuum tunneling. What if we still live in the wrong vacuum? This is quite possible, if today the cosmological constant is not truly zero. Then a bubble of true vacuum may nucleate at any

time anywhere in the Universe. Such a bubble will grow with the velocity of light, turning the whole Universe into the true vacuum. For us this would be disastrous, because nothing in the old Universe will survive this change. Also such an event is impossible to predict, and could basically happen tomorrow. I guess we should all feel much sadder, when some clever string theorist finally proves that the true vacuum is $SU(3) \otimes SU(2) \otimes U(1)$ invariant.

Much of this last section is speculative, and — in any case — not testable in observations. Many of the above ideas may have to be revised, as our knowledge increases. Still, whatever we may learn in the future, it is likely that the Universe will remain an interesting and lively place for a very long time. We can only hope that the same is true for our own earth.

Appendix A

Quantum Field Theory

This chapter provides a brief summary of the mathematical structure of quantum field theory. Classical field theories are discussed as a generalization of point mechanics to systems with infinitely many degrees of freedom — a given number per space point. Similarly, quantum field theories are just quantum mechanical systems with infinitely many degrees of freedom. In the same way as point mechanics systems, classical field theories can be quantized with path integral methods. The quantization of field theories at finite temperature leads to path integrals in Euclidean time. This provides us with an analogy between quantum field theory and classical statistical mechanics. We also mention the lattice regularization which has recently provided a mathematically satisfactory formulation of the standard model beyond perturbation theory.

A.1 From Point Mechanics to Classical Field Theory

Point mechanics describes the dynamics of classical nonrelativistic point particles. The coordinates of the particles represent a finite number of degrees of freedom. In the simplest case — a single particle moving in one spatial dimension — we are dealing with a single degree of freedom: the x -coordinate of the particle. The dynamics of a particle of mass m moving in an external potential $V(x)$ is described by Newton's equation

$$m\partial_t^2 x = ma = F(x) = -\frac{dV(x)}{dx}. \quad (\text{A.1.1})$$

Once the initial conditions are specified, this ordinary second order differential equation determines the particle's path $x(t)$, i.e. its position as a function of time. Newton's equation results from the variational principle to minimize the action

$$S[x] = \int dt L(x, \partial_t x), \quad (\text{A.1.2})$$

over the space of all paths $x(t)$. The action is a functional (a function whose argument is itself a function) that results from the time integral of the Lagrange function

$$L(x, \partial_t x) = \frac{m}{2}(\partial_t x)^2 - V(x). \quad (\text{A.1.3})$$

The Euler-Lagrange equation

$$\partial_t \frac{\delta L}{\delta(\partial_t x)} - \frac{\delta L}{\delta x} = 0, \quad (\text{A.1.4})$$

is nothing but Newton's equation.

Classical field theories are a generalization of point mechanics to systems with infinitely many degrees of freedom — a given number for each space point \vec{x} . In this case, the degrees of freedom are the field values $\Phi(\vec{x})$, where Φ is some generic field. In case of a neutral scalar field, Φ is simply a real number representing one degree of freedom per space point. A charged scalar field, on the other hand, is described by a complex number and hence represents two degrees of freedom per space point. The scalar Higgs field $\Phi^a(\vec{x})$ (with $a \in \{1, 2\}$) in the standard model is a complex doublet, i.e. it has four real degrees of freedom per space point. An Abelian gauge field $A_i(\vec{x})$ (with a spatial direction index $i \in \{1, 2, 3\}$) — for example, the photon field in electrodynamics — is a neutral vector field with 3 real degrees of freedom per space point. One of these degrees of freedom is redundant due to the $U(1)$ gauge symmetry. Hence, an Abelian gauge field has two physical degrees of freedom per space point which correspond to the two polarization states of the massless photon. Note that the time-component $A_0(\vec{x})$ does not represent a physical degree of freedom. It is just a Lagrange multiplier field that enforces the Gauss law. A non-Abelian gauge field $A_i^a(\vec{x})$ is charged and has an additional index a . For example, the gluon field in chromodynamics with a color index $a \in \{1, 2, \dots, 8\}$ represents $2 \times 8 = 16$ physical degrees of freedom per space point, again because of some redundancy due to the $SU(3)_c$ color gauge symmetry. The field that represents the W - and Z -bosons in the standard model has an index $a \in \{1, 2, 3\}$ and transforms under the gauge group $SU(2)_L$. Thus, it represents $2 \times 3 = 6$ physical degrees of freedom. However, in contrast to the photon, the W - and Z -bosons are massive due to the Higgs mechanism and have three (not just two) polarization states. The extra degree of freedom is provided by the Higgs field.

The analog of Newton's equation in field theory is the classical field equation of motion. For example, for a neutral scalar field this is the Klein-Gordon equation

$$\partial^\mu \partial_\mu \Phi = -\frac{dV(\Phi)}{d\Phi}. \quad (\text{A.1.5})$$

Again, after specifying appropriate initial conditions it determines the classical field configuration $\Phi(x)$, i.e. the values of the field Φ at all space-time points $x = (t, \vec{x})$. Hence, the role of time in point mechanics is played by space-time in field theory, and the role of the point particle coordinates is now played by the field values. As before, the classical equation of motion results from minimizing the action

$$S[\Phi] = \int d^4x \mathcal{L}(\Phi, \partial_\mu \Phi). \quad (\text{A.1.6})$$

The integral over time in eq.(A.1.2) is now replaced by an integral over space-time and the Lagrange function of point mechanics gets replaced by the Lagrange density function (or Lagrangian)

$$\mathcal{L}(\Phi, \partial_\mu \Phi) = \frac{1}{2} \partial^\mu \Phi \partial_\mu \Phi - V(\Phi). \quad (\text{A.1.7})$$

A simple interacting field theory is the Φ^4 theory with the potential

$$V(\Phi) = \frac{m^2}{2} \Phi^2 + \frac{\lambda}{4} \Phi^4. \quad (\text{A.1.8})$$

Here m is the mass of the scalar field and λ is the coupling strength of its self-interaction. Note that the mass term corresponds to a harmonic oscillator potential in the point mechanics analog, while the interaction term corresponds to an anharmonic perturbation. As before, the Euler-Lagrange equation

$$\partial_\mu \frac{\delta L}{\delta(\partial_\mu \Phi)} - \frac{\delta L}{\delta \Phi} = 0, \quad (\text{A.1.9})$$

is the classical equation of motion, in this case the Klein-Gordon equation. The analogies between point mechanics and field theory are summarized in table A.1.

A.2 Path Integral in Real Time

The quantization of field theories is most conveniently performed using the path integral approach. Here we first discuss the path integral in quantum mechanics

Point Mechanics	Field Theory
time t	space-time $x = (t, \vec{x})$
particle coordinate x	field value Φ
particle path $x(t)$	field configuration $\Phi(x)$
action $S[x] = \int dt L(x, \partial_t x)$	action $S[\Phi] = \int d^4x \mathcal{L}(\Phi, \partial_\mu \Phi)$
Lagrange function $L(x, \partial_t x) = \frac{m}{2}(\partial_t x)^2 - V(x)$	Lagrangian $\mathcal{L}(\Phi, \partial_\mu \Phi) = \frac{1}{2}\partial^\mu \Phi \partial_\mu \Phi - V(\Phi)$
equation of motion $\partial_t \frac{\delta L}{\delta(\partial_t x)} - \frac{\delta L}{\delta x} = 0$	field equation $\partial_\mu \frac{\delta \mathcal{L}}{\delta(\partial_\mu \Phi)} - \frac{\delta \mathcal{L}}{\delta \Phi} = 0$
Newton's equation $\partial_t^2 x = -\frac{dV(x)}{dx}$	Klein-Gordon equation $\partial^\mu \partial_\mu \Phi = -\frac{dV(\Phi)}{d\Phi}$
kinetic energy $\frac{m}{2}(\partial_t x)^2$	kinetic energy $\frac{1}{2}\partial^\mu \Phi \partial_\mu \Phi$
harmonic oscillator potential $\frac{m}{2}\omega^2 x^2$	mass term $\frac{m^2}{2}\Phi^2$
anharmonic perturbation $\frac{\lambda}{4}x^4$	self-interaction term $\frac{\lambda}{4}\Phi^4$

Table A.1: *The dictionary that translates point mechanics into the language of field theory.*

— quantized point mechanics — using the real time formalism. A mathematically more satisfactory formulation uses an analytic continuation to so-called Euclidean time. This will be discussed in the next section.

The real time evolution of a quantum system described by a Hamilton operator H is given by the time-dependent Schrödinger equation

$$i\hbar \partial_t |\Psi(t)\rangle = H |\Psi(t)\rangle. \quad (\text{A.2.1})$$

For a time-independent Hamilton operator the time evolution operator is given by

$$U(t', t) = \exp\left(-\frac{i}{\hbar} H(t' - t)\right), \quad (\text{A.2.2})$$

such that

$$|\Psi(t')\rangle = U(t', t) |\Psi(t)\rangle. \quad (\text{A.2.3})$$

Let us consider the transition amplitude $\langle x' | U(t', t) | x \rangle$ of a nonrelativistic point particle that starts at position x at time t and arrives at position x' at time t' . Using

$$\langle x | \Psi(t) \rangle = \Psi(x, t) \quad (\text{A.2.4})$$

we obtain

$$\Psi(x', t') = \int dx \langle x' | U(t', t) | x \rangle \Psi(x, t), \quad (\text{A.2.5})$$

i.e. $\langle x'|U(t', t)|x\rangle$ acts as a propagator for the wave function. The propagator is of physical interest because it contains information about the energy spectrum. When we consider propagation from an initial position x back to the same position we find

$$\begin{aligned}\langle x|U(t', t)|x\rangle &= \langle x|\exp(-\frac{i}{\hbar}H(t' - t))|x\rangle \\ &= \sum_n |\langle x|n\rangle|^2 \exp(-\frac{i}{\hbar}E_n(t' - t)).\end{aligned}\quad (\text{A.2.6})$$

We have inserted a complete set, $\sum_n |n\rangle\langle n| = \mathbb{1}$, of energy eigenstates $|n\rangle$ with

$$H|n\rangle = E_n|n\rangle. \quad (\text{A.2.7})$$

Hence, according to eq.(A.2.6), the Fourier transform of the propagator yields the energy spectrum as well as the energy eigenstates $\langle x|n\rangle$.

Inserting a complete set of position eigenstates we arrive at

$$\begin{aligned}\langle x'|U(t', t)|x\rangle &= \langle x'|\exp(-\frac{i}{\hbar}H(t' - t_1 + t_1 - t))|x\rangle \\ &= \int dx_1 \langle x'| \exp(-\frac{i}{\hbar}H(t' - t_1))|x_1\rangle \\ &\times \langle x_1| \exp(-\frac{i}{\hbar}H(t_1 - t))|x\rangle \\ &= \int dx_1 \langle x'|U(t', t_1)|x_1\rangle \langle x_1|U(t_1, t)|x\rangle.\end{aligned}\quad (\text{A.2.8})$$

It is obvious that we can repeat this process an arbitrary number of times. This is exactly what we do in the formulation of the path integral. Let us divide the time interval $[t, t']$ into N elementary time steps of size ε such that

$$t' - t = N\varepsilon. \quad (\text{A.2.9})$$

Inserting a complete set of position eigenstates at the intermediate times $t_i, i \in \{1, 2, \dots, N-1\}$ we obtain

$$\begin{aligned}\langle x'|U(t', t)|x\rangle &= \int dx_1 \int dx_2 \dots \int dx_{N-1} \langle x'|U(t', t_{N-1})|x_{N-1}\rangle \dots \\ &\times \langle x_2|U(t_2, t_1)|x_1\rangle \langle x_1|U(t_1, t)|x\rangle.\end{aligned}\quad (\text{A.2.10})$$

In the next step we concentrate on one of the factors and we consider a single nonrelativistic point particle moving in an external potential $V(x)$ such that

$$H = \frac{p^2}{2m} + V(x). \quad (\text{A.2.11})$$

Using the Baker-Campbell-Hausdorff formula and neglecting terms of order ε^2 we find

$$\begin{aligned}
\langle x_{i+1}|U(t_{i+1}, t_i)|x_i\rangle &= \langle x_{i+1}|\exp(-\frac{i\varepsilon p^2}{2m\hbar})\exp(-\frac{i\varepsilon}{\hbar}V(x))|x_i\rangle \\
&= \frac{1}{2\pi} \int dp \langle x_{i+1}|\exp(-\frac{i\varepsilon p^2}{2m\hbar})|p\rangle \langle p|\exp(-\frac{i\varepsilon}{\hbar}V(x))|x_i\rangle \\
&= \frac{1}{2\pi} \int dp \exp(-\frac{i\varepsilon p^2}{2m\hbar}) \exp(-\frac{i}{\hbar}p(x_{i+1} - x_i)) \\
&\times \exp(-\frac{i\varepsilon}{\hbar}V(x_i)). \tag{A.2.12}
\end{aligned}$$

The integral over p is ill-defined because the integrand is a very rapidly oscillating function. To make the expression well-defined we replace the time step ε by $\varepsilon - ia$, i.e. we go out into the complex time plane. After doing the integral we take the limit $a \rightarrow 0$. Still, one should keep in mind that the definition of the path integral required an analytic continuation in time. One finds

$$\langle x_{i+1}|U(t_{i+1}, t_i)|x_i\rangle = \sqrt{\frac{m}{2\pi i\hbar\varepsilon}} \exp(\frac{i}{\hbar}\varepsilon[\frac{m}{2}(\frac{x_{i+1} - x_i}{\varepsilon})^2 - V(x_i)]). \tag{A.2.13}$$

Inserting this back into the expression for the propagator we obtain

$$\langle x'|U(t', t)|x\rangle = \int \mathcal{D}x \exp(\frac{i}{\hbar}S[x]). \tag{A.2.14}$$

The action has been identified in the time continuum limit as

$$\begin{aligned}
S[x] &= \int dt [\frac{m}{2}(\partial_t x)^2 - V(x)] \\
&= \lim_{\varepsilon \rightarrow 0} \sum_i \varepsilon [\frac{m}{2}(\frac{x_{i+1} - x_i}{\varepsilon})^2 - V(x_i)]. \tag{A.2.15}
\end{aligned}$$

The integration measure is defined as

$$\int \mathcal{D}x = \lim_{\varepsilon \rightarrow 0} \sqrt{\frac{m}{2\pi i\hbar\varepsilon}}^{N-1} \int dx_1 \int dx_2 \dots \int dx_{N-1}. \tag{A.2.16}$$

This means that we integrate over the possible particle positions for each intermediate time t_i . In this way we integrate over all possible paths of the particle starting at x and ending at x' . Each path is weighted with an oscillating phase factor $\exp(\frac{i}{\hbar}S[x])$ determined by the action. The classical path of minimum action has the smallest oscillations, and hence the largest contribution to the path integral. In the classical limit $\hbar \rightarrow 0$ only that contribution survives.

A.3 Path Integral in Euclidean Time

As we have seen, it requires a small excursion into the complex time plane to make the path integral mathematically well-defined. Now we will make a big step into that plane and actually consider purely imaginary so-called Euclidean time. The physical motivation for this, however, comes from quantum statistical mechanics. Let us consider the quantum statistical partition function

$$Z = \text{Tr} \exp(-\beta H), \quad (\text{A.3.1})$$

where $\beta = 1/T$ is the inverse temperature. It is mathematically equivalent to the time interval we discussed in the real time path integral. In particular, the operator $\exp(-\beta H)$ turns into the time evolution operator $U(t', t)$ if we identify

$$\beta = \frac{i}{\hbar}(t' - t). \quad (\text{A.3.2})$$

In this sense the system at finite temperature corresponds to a system propagating in purely imaginary (Euclidean) time. By dividing the Euclidean time interval into N time steps, i.e. by writing $\beta = Na/\hbar$, and again by inserting complete sets of position eigenstates we now arrive at the Euclidean time path integral

$$Z = \int \mathcal{D}x \exp(-\frac{1}{\hbar} S_E[x]). \quad (\text{A.3.3})$$

The action now takes the Euclidean form

$$\begin{aligned} S_E[x] &= \int dt \left[\frac{m}{2} (\partial_t x)^2 + V(x) \right] \\ &= \lim_{a \rightarrow 0} \sum_i a \left[\frac{m}{2} \left(\frac{x_{i+1} - x_i}{a} \right)^2 + V(x_i) \right]. \end{aligned} \quad (\text{A.3.4})$$

In contrast to the real time case the measure now involves N integrations

$$\int \mathcal{D}x = \lim_{a \rightarrow 0} \sqrt{\frac{m}{2\pi\hbar a}}^N \int dx_1 \int dx_2 \dots \int dx_N. \quad (\text{A.3.5})$$

The extra integration over $x_N = x'$ is due to the trace in eq.(A.3.1). Note that there is no extra integration over $x_0 = x$ because the trace implies periodic boundary conditions in the Euclidean time direction, i.e. $x_0 = x_N$.

The Euclidean path integral allows us to evaluate thermal expectation values. For example, let us consider an operator $\mathcal{O}(x)$ that is diagonal in the position

state basis. We can insert this operator in the path integral and thus compute its expectation value

$$\langle \mathcal{O}(x) \rangle = \frac{1}{Z} \text{Tr}[\mathcal{O}(x) \exp(-\beta H)] = \frac{1}{Z} \int \mathcal{D}x \, \mathcal{O}(x(0)) \exp(-\frac{1}{\hbar} S_E[x]). \quad (\text{A.3.6})$$

Since the theory is translation invariant in Euclidean time one can place the operator anywhere in time, e.g. at $t = 0$ as done here. When we perform the low temperature limit, $\beta \rightarrow \infty$, the thermal fluctuations are switched off and only the quantum ground state $|0\rangle$ (the vacuum) contributes to the partition function, i.e. $Z \sim \exp(-\beta E_0)$. In this limit the path integral is formulated in an infinite Euclidean time interval, and describes the vacuum expectation value

$$\langle \mathcal{O}(x) \rangle = \langle 0 | \mathcal{O}(x) | 0 \rangle = \lim_{\beta \rightarrow \infty} \frac{1}{Z} \int \mathcal{D}x \, \mathcal{O}(x(0)) \exp(-\frac{1}{\hbar} S_E[x]). \quad (\text{A.3.7})$$

It is also interesting to consider 2-point functions of operators at different instances in Euclidean time

$$\begin{aligned} \langle \mathcal{O}(x(0)) \mathcal{O}(x(t)) \rangle &= \frac{1}{Z} \text{Tr}[\mathcal{O}(x) \exp(-Ht) \mathcal{O}(x) \exp(Ht) \exp(-\beta H)] \\ &= \frac{1}{Z} \int \mathcal{D}x \, \mathcal{O}(x(0)) \mathcal{O}(x(t)) \exp(-\frac{1}{\hbar} S_E[x]). \end{aligned} \quad (\text{A.3.8})$$

Again, we consider the limit $\beta \rightarrow \infty$, but we also separate the operators in time, i.e. we also let $t \rightarrow \infty$. Then the leading contribution is $|\langle 0 | \mathcal{O}(x) | 0 \rangle|^2$. Subtracting this, and thus forming the connected 2-point function, one obtains

$$\lim_{\beta, t \rightarrow \infty} \langle \mathcal{O}(x(0)) \mathcal{O}(x(t)) \rangle - |\langle \mathcal{O}(x) \rangle|^2 = |\langle 0 | \mathcal{O}(x) | 1 \rangle|^2 \exp(-(E_1 - E_0)t). \quad (\text{A.3.9})$$

Here $|1\rangle$ is the first excited state of the quantum system with an energy E_1 . The connected 2-point function decays exponentially at large Euclidean time separations. The decay is governed by the energy gap $E_1 - E_0$. In a quantum field theory E_1 corresponds to the energy of the lightest particle. Its mass is determined by the energy gap $E_1 - E_0$ above the vacuum. Hence, in Euclidean field theory particle masses are determined from the exponential decay of connected 2-point correlation functions.

A.4 Spin Models in Classical Statistical Mechanics

So far we have considered quantum systems both at zero and at finite temperature. We have represented their partition functions as Euclidean path integrals

over configurations on a time lattice of length β . We will now make a completely new start and study classical discrete systems at finite temperature. We will see that their mathematical description is very similar to the path integral formulation of quantum systems. Still, the physical interpretation of the formalism is drastically different in the two cases. In the next section we will set up a dictionary that allows us to translate quantum physics language into the language of classical statistical mechanics.

For simplicity, let us concentrate on simple classical spin models. Here the word spin does not mean that we deal with quantized angular momenta. All we do is work with classical variables that can point in specific directions. The simplest spin model is the Ising model with classical spin variables $s_x = \pm 1$. (Again, these do not represent the quantum states up and down of a quantum mechanical angular momentum $1/2$.) More complicated spin models with an $O(N)$ spin rotational symmetry are the XY model ($N = 2$) and the Heisenberg model ($N = 3$). The spins in the XY model are 2-component unit-vectors, while the spins in the Heisenberg model have three components. In all these models the spins live on the sites of a d -dimensional spatial lattice. The lattice is meant to be a crystal lattice (so typically $d = 3$) and the lattice spacing has a physical meaning. This is in contrast to the Euclidean time lattice that we have introduced to make the path integral mathematically well-defined, and that we finally send to zero in order to reach the Euclidean time continuum limit. The Ising model is characterized by its classical Hamilton function (not a quantum Hamilton operator) which simply specifies the energy of any configuration of spins. The Ising Hamilton function is a sum of nearest neighbor contributions

$$\mathcal{H}[s] = J \sum_{\langle xy \rangle} s_x s_y - \mu B \sum_x s_x, \quad (\text{A.4.1})$$

with a ferromagnetic coupling constant $J < 0$ that favors parallel spins, plus a coupling to an external magnetic field B . The classical partition function of this system is given by

$$Z = \int \mathcal{D}s \exp(-\mathcal{H}[s]/T) = \prod_x \sum_{s_x = \pm 1} \exp(-\mathcal{H}[s]/T). \quad (\text{A.4.2})$$

The sum over all spin configurations corresponds to an independent summation over all possible orientations of individual spins. Thermal averages are computed by inserting appropriate operators. For example, the magnetization is given by

$$\langle s_x \rangle = \frac{1}{Z} \prod_x \sum_{s_x = \pm 1} s_x \exp(-\mathcal{H}[s]/T). \quad (\text{A.4.3})$$

Similarly, the spin correlation function is defined by

$$\langle s_x s_y \rangle = \frac{1}{Z} \prod_x \sum_{s_x = \pm 1} s_x s_y \exp(-\mathcal{H}[s]/T). \quad (\text{A.4.4})$$

At large distances the connected spin correlation function typically decays exponentially

$$\langle s_x s_y \rangle - \langle s \rangle^2 \sim \exp(-|x - y|/\xi), \quad (\text{A.4.5})$$

where ξ is the so-called correlation length. At general temperatures the correlation length is typically just a few lattice spacings. When one models real materials, the Ising model would generally be a great oversimplification, because real magnets, for example, not only have nearest neighbor couplings. Still, the details of the Hamilton function at the scale of the lattice spacing are not always important. There is a critical temperature T_c at which ξ diverges and universal behavior arises. At this temperature a second order phase transition occurs. Then the details of the model at the scale of the lattice spacing are irrelevant for the long range physics that takes place at the scale of ξ . In fact, at their critical temperatures some real materials behave just like the simple Ising model. This is why the Ising model is so interesting. It is just a very simple member of a large universality class of different models, which all share the same critical behavior. This does not mean that they have the same values of their critical temperatures. However, their magnetization goes to zero at the critical temperature with the same power of $T_c - T$, i.e. their critical exponents are identical.

A.5 Quantum Mechanics versus Statistical Mechanics

We notice a close analogy between the Euclidean path integral for a quantum mechanical system and a classical statistical mechanics system like the Ising model. The path integral for the quantum system is defined on a 1-dimensional Euclidean time lattice, just like an Ising model can be defined on a d -dimensional spatial lattice. In the path integral we integrate over all paths, i.e. over all configurations $x(t)$, while in the Ising model we sum over all spin configurations s_x . Paths are weighted by their Euclidean action $S_E[x]$ while spin configurations are weighted with their Boltzmann factors depending on the classical Hamilton function $\mathcal{H}[s]$. The prefactor of the action is $1/\hbar$, and the prefactor of the Hamilton function is $1/T$. Indeed \hbar determines the strength of quantum fluctuations, while the temperature T determines the strength of thermal fluctuations. The kinetic energy $\frac{1}{2}((x_{i+1} - x_i)/a)^2$ in the path integral is analogous to the nearest neighbor spin coupling $s_x s_{x+1}$, and the potential term $V(x_i)$ is analogous to the coupling $\mu B s_x$.

Quantum mechanics	Classical statistical mechanics
Euclidean time lattice	d -dimensional spatial lattice
elementary time step a	crystal lattice spacing
particle position x	classical spin variable s
particle path $x(t)$	spin configuration s_x
path integral $\int \mathcal{D}x$	sum over configurations $\prod_x \sum_{s_x}$
Euclidean action $S_E[x]$	classical Hamilton function $\mathcal{H}[s]$
Planck's constant \hbar	temperature T
quantum fluctuations	thermal fluctuations
kinetic energy $\frac{1}{2}(\frac{x_{i+1}-x_i}{a})^2$	neighbor coupling $s_x s_{x+1}$
potential energy $V(x_i)$	external field energy $\mu B s_x$
weight of a path $\exp(-\frac{1}{\hbar} S_E[x])$	Boltzmann factor $\exp(-\mathcal{H}[s]/T)$
vacuum expectation value $\langle \mathcal{O}(x) \rangle$	magnetization $\langle s_x \rangle$
2-point function $\langle \mathcal{O}(x(0)) \mathcal{O}(x(t)) \rangle$	correlation function $\langle s_x s_y \rangle$
energy gap $E_1 - E_0$	inverse correlation length $1/\xi$
continuum limit $a \rightarrow 0$	critical behavior $\xi \rightarrow \infty$

Table A.2: *The dictionary that translates quantum mechanics into the language of classical statistical mechanics.*

to an external magnetic field. The magnetization $\langle s_x \rangle$ corresponds to the vacuum expectation value of an operator $\langle \mathcal{O}(x) \rangle$ and the spin-spin correlation function $\langle s_x s_y \rangle$ corresponds to the 2-point correlation function $\langle \mathcal{O}(x(0)) \mathcal{O}(x(t)) \rangle$. The inverse correlation length $1/\xi$ is analogous to the energy gap $E_1 - E_0$ (and hence to a particle mass in a Euclidean quantum field theory). Finally, the Euclidean time continuum limit $a \rightarrow 0$ corresponds to a second order phase transition where $\xi \rightarrow \infty$. The lattice spacing in the path integral is an artifact of our mathematical description which we send to zero while the physics remains constant. In classical statistical mechanics, on the other hand, the lattice spacing is physical and hence fixed, while the correlation length ξ goes to infinity at a second order phase transition. All this is summarized in the dictionary of table A.2.

A.6 Lattice Field Theory

So far we have restricted ourselves to quantum mechanical problems and to classical statistical mechanics. The former were defined by a path integral on a 1-d Euclidean time lattice, while the latter involved spin models on a d -dimensional

spatial lattice. When we quantize field theories on the lattice, we formulate the theory on a d -dimensional space-time lattice, i.e. usually the lattice is 4-dimensional. Just as we integrate over all configurations (all paths) $x(t)$ of a quantum particle, we now integrate over all configurations $\Phi(x)$ of a quantum field defined at any Euclidean space-time point $x = (\vec{x}, x_4)$. Again the weight factor in the path integral is given by the action. Let us illustrate this for a free neutral scalar field $\Phi(x) \in R$. Its Euclidean action is given by

$$S_E[\Phi] = \int d^4x \left[\frac{1}{2} \partial_\mu \Phi \partial_\mu \Phi + \frac{m^2}{2} \Phi^2 \right]. \quad (\text{A.6.1})$$

Interactions can be included, for example, by adding a $\frac{\lambda}{4} \Phi^4$ term to the action. The Feynman path integral for this system is formally written as

$$Z = \int \mathcal{D}\Phi \exp(-S_E[\Phi]). \quad (\text{A.6.2})$$

(Note that we have put $\hbar = c = 1$.) The integral is over all field configurations, which is a divergent expression if no regularization is imposed. One can make the expression mathematically well-defined by using dimensional regularization of Feynman diagrams. This approach is, however, limited to perturbation theory. The lattice allows us to formulate field theory beyond perturbation theory, which is very essential for strongly interacting theories like QCD, but also for the standard model in general. For example, due to the heavy mass of the top quark, the Yukawa coupling between the Higgs and top quark field is rather strong. The above free scalar field theory, of course, does not really require a nonperturbative treatment. We use it only to illustrate the lattice quantization method in a simple setting. On the lattice the continuum field $\Phi(x)$ is replaced by a lattice field Φ_x , which is restricted to the points x of a d -dimensional space-time lattice. From now on we will work in lattice units, i.e. we put $a = 1$. The above continuum action can be approximated by discretizing the continuum derivatives such that

$$S_E[\Phi] = \sum_{x,\mu} \frac{1}{2} (\Phi_{x+\hat{\mu}} - \Phi_x)^2 + \sum_x \frac{m^2}{2} \Phi_x^2. \quad (\text{A.6.3})$$

Here $\hat{\mu}$ is the unit vector in the μ -direction. The integral over all field configurations now becomes a multiple integral over all values of the field at all lattice points

$$Z = \prod_x \int_{-\infty}^{\infty} d\Phi_x \exp(-S_E[\Phi]). \quad (\text{A.6.4})$$

For a free field theory the partition function is just a Gaussian integral. In fact, one can write the lattice action as

$$S_E[\Phi] = \frac{1}{2} \sum_{x,y} \Phi_x \mathcal{M}_{xy} \Phi_y, \quad (\text{A.6.5})$$

where the matrix \mathcal{M} describes the couplings between lattice points. Diagonalizing this matrix by an orthogonal transformation \mathcal{O} one has

$$\mathcal{M} = \mathcal{O}^T \mathcal{D} \mathcal{O}. \quad (\text{A.6.6})$$

Introducing

$$\Phi'_x = \mathcal{O}_{xy} \Phi_y \quad (\text{A.6.7})$$

one obtains

$$Z = \prod_x \int d\Phi'_x \exp\left(-\frac{1}{2} \sum_x \Phi'_x \mathcal{D}_{xx} \Phi'_x\right) = (2\pi)^{N/2} \det \mathcal{D}^{-1/2}, \quad (\text{A.6.8})$$

where N is the number of lattice points.

To extract the energy values of the corresponding quantum Hamilton operator we need to study the 2-point function of the lattice field

$$\langle \Phi_x \Phi_y \rangle = \frac{1}{Z} \int \mathcal{D}\Phi \Phi_x \Phi_y \exp(-S_E[\Phi]). \quad (\text{A.6.9})$$

This is most conveniently done by introducing a source field in the partition function, such that

$$Z[J] = \int \mathcal{D}\Phi \exp(-S_E[\Phi] + \sum_x J_x \Phi_x). \quad (\text{A.6.10})$$

Then the connected 2-point function is given by

$$\langle \Phi_x \Phi_y \rangle - \langle \Phi \rangle^2 = \frac{\partial^2 \log Z[J]}{\partial J_x \partial J_y} \Big|_{J=0}. \quad (\text{A.6.11})$$

The Boltzmann factor characterizing the problem with the external sources is given by the exponent

$$\frac{1}{2} \Phi \mathcal{M} \Phi - J \Phi = \frac{1}{2} \Phi' \mathcal{M} \Phi' - \frac{1}{2} J \mathcal{M}^{-1} J. \quad (\text{A.6.12})$$

Here we have introduced

$$\Phi' = \Phi - \mathcal{M}^{-1} J. \quad (\text{A.6.13})$$

Integrating over Φ' in the path integral we obtain

$$Z[J] = (2\pi)^{N/2} \det \mathcal{D}^{-1/2} \exp\left(\frac{1}{2} J \mathcal{M}^{-1} J\right), \quad (\text{A.6.14})$$

and hence

$$\langle \Phi_x \Phi_y \rangle = \frac{1}{2} \mathcal{M}_{xy}^{-1}. \quad (\text{A.6.15})$$

It is instructive to invert the matrix \mathcal{M} by going to Fourier space, i.e. by writing

$$\Phi_x = \frac{1}{(2\pi)^d} \int_B d^d p \Phi(p) \exp(ipx). \quad (\text{A.6.16})$$

The momentum space of the lattice is given by the Brillouin zone $B =]-\pi, \pi]^d$. For the 2-point function in momentum space one then finds

$$\langle \Phi(-p)\Phi(p) \rangle = [\sum_{\mu} (2 \sin(p_{\mu}/2))^2 + m^2]^{-1}. \quad (\text{A.6.17})$$

This is the lattice version of the continuum propagator

$$\langle \Phi(-p)\Phi(p) \rangle = (p^2 + m^2)^{-1}. \quad (\text{A.6.18})$$

From the lattice propagator we can deduce the energy spectrum of the lattice theory. For this purpose we construct a lattice field with definite spatial momentum \vec{p} located in a specific time slice

$$\Phi(\vec{p})_t = \sum_x \Phi_{\vec{x},t} \exp(-i\vec{p} \cdot \vec{x}), \quad (\text{A.6.19})$$

and we consider its 2-point function

$$\langle \Phi(-\vec{p})_0 \Phi(\vec{p})_t \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} dp_d \langle \Phi(-p)\Phi(p) \rangle \exp(ip_d t). \quad (\text{A.6.20})$$

Inserting the lattice propagator of eq.(A.6.17) one can perform the integral. One encounters a pole in the propagator when $p_d = iE$ with

$$(2 \sinh(E/2))^2 = \sum_i (2 \sin(p_i/2))^2 + m^2. \quad (\text{A.6.21})$$

The 2-point function then takes the form

$$\langle \Phi(-\vec{p})_0 \Phi(\vec{p})_t \rangle = C \exp(-Et), \quad (\text{A.6.22})$$

i.e. it decays exponentially with slope E . This allows us to identify E as the energy of the lattice scalar particle with spatial momentum \vec{p} . In general, E differs from the correct continuum dispersion relation

$$E^2 = \vec{p}^2 + m^2. \quad (\text{A.6.23})$$

Only in the continuum limit, i.e. when E , \vec{p} and m are small in lattice units, the lattice dispersion relation agrees with the one of the continuum theory.

We have defined the path integral by using the action of the classical theory. Theories with fermions have no immediate classical limit, and the definition of the path integral needs special care. The first step is to define a so-called Grassmann algebra, which works with anticommuting classical variables η_i with $i \in 1, 2, \dots, N$. A Grassmann algebra is characterized by the anticommutation relations

$$\{\eta_i, \eta_j\} = \eta_i \eta_j + \eta_j \eta_i = 0. \quad (\text{A.6.24})$$

An element of the Grassmann algebra is a polynomial in the generators

$$f(\eta) = f + \sum_i f_i \eta_i + \sum_{ij} f_{ij} \eta_i \eta_j + \sum_{ijk} f_{ijk} \eta_i \eta_j \eta_k + \dots \quad (\text{A.6.25})$$

The $f_{ij\dots l}$ are ordinary complex numbers antisymmetric in i, j, \dots, l . One also defines formal differentiation and integration procedures. The differentiation rules are

$$\frac{\partial}{\partial \eta_i} \eta_i = 1, \quad \frac{\partial}{\partial \eta_i} \eta_i \eta_j = \eta_j, \quad \frac{\partial}{\partial \eta_i} \eta_j \eta_i = -\eta_j, \quad (\text{A.6.26})$$

and integration is defined by

$$\int d\eta_i = 0, \quad \int d\eta_i \eta_i = 1, \quad \int d\eta_i d\eta_j \eta_i \eta_j = -1. \quad (\text{A.6.27})$$

These integrals are formal expressions. It has no meaning to ask over which range of η_i values we actually integrate.

The Grassmann algebra we use to define fermion fields is generated by Grassmann numbers Ψ_x and $\bar{\Psi}_x$, which are completely independent. The index x runs over all space-time points as well as over all spin, flavor, color or other indices. Let us consider the simplest (completely unrealistic) case of just two degrees of freedom Ψ and $\bar{\Psi}$, and let us perform the Gaussian integral

$$\int d\bar{\Psi} d\Psi \exp(-m\bar{\Psi}\Psi) = \int d\bar{\Psi} d\Psi (1 - m\bar{\Psi}\Psi) = m. \quad (\text{A.6.28})$$

Note that the expansion of the exponential terminates because $\Psi^2 = \bar{\Psi}^2 = 0$. When we enlarge the Grassmann algebra to an arbitrary number of elements the above formula generalizes to

$$\prod_x \int d\bar{\Psi}_x d\Psi_x \exp(-\bar{\Psi}_x \mathcal{M}_{xy} \Psi_y) = \int \mathcal{D}\bar{\Psi} \mathcal{D}\Psi \exp(-\bar{\Psi} \mathcal{M} \Psi) = \det \mathcal{M}. \quad (\text{A.6.29})$$

In the two variable case we have

$$\int d\bar{\Psi} d\Psi \bar{\Psi} \Psi \exp(-m\bar{\Psi}\Psi) = 1, \quad (\text{A.6.30})$$

which generalizes to

$$\int \mathcal{D}\bar{\Psi}\mathcal{D}\Psi \bar{\Psi}_x\Psi_y \exp(-\bar{\Psi}\mathcal{M}\Psi_y) = \mathcal{M}_{ij}^{-1}\det\mathcal{M}. \quad (\text{A.6.31})$$

Lattice fermions have several technical problems that have prevented the non-perturbative formulation of the standard model for many years. For example, chiral fermions — like neutrinos — suffer from the lattice fermion doubling problem. Every left-handed neutrino necessarily comes with a right-handed partner. Until recently, it was not known how to couple only the left-handed particles to an electroweak lattice gauge field. Thanks to a recent breakthrough in lattice gauge theory, the standard model is now consistently defined beyond perturbation theory. Even the perturbative definition of the standard model has been incomplete beyond one loop, due to ambiguities in treating γ_5 in dimensional regularization. All these ambiguities are now eliminated, thanks to the new lattice result. In the rest of this course, we will not need the lattice regularization much further. Still, it is good to know that the standard model now stands on a firm mathematical basis and that the path integral expressions we write down for it are completely well-defined even beyond perturbation theory.

Appendix B

Group Theory of S_N and $SU(n)$

B.1 The Permutation Group S_N

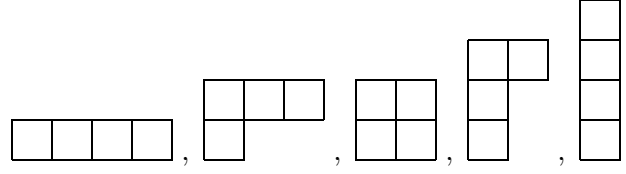
Let us consider the permutation symmetry of N objects — for example the fundamental representations of $SU(n)$. Their permutations form the group S_N . The permutation group has $N!$ elements — all permutations of N objects. The group S_2 has two elements: the identity and the pair permutation. The representations of S_2 are represented by Young tableaux

$$\begin{array}{ll}
 \begin{array}{|c|c|} \hline \square & \square \\ \hline \end{array} & \text{1-dimensional symmetric representation,} \\
 \begin{array}{|c|} \hline \square \\ \hline \square \\ \hline \end{array} & \text{1-dimensional antisymmetric representation.}
 \end{array} \tag{B.1.1}$$

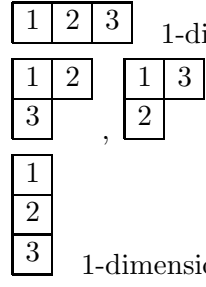
To describe the permutation properties of three objects we need the group S_3 . It has $3! = 6$ elements: the identity, 3 pair permutations and 2 cyclic permutations. The group S_3 has three irreducible representations

$$\begin{array}{ll}
 \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \end{array} & \text{1-dimensional symmetric representation,} \\
 \begin{array}{|c|c|} \hline \square & \square \\ \hline \square & \\ \hline \end{array} & \text{2-dimensional representation of mixed symmetry,} \\
 \begin{array}{|c|} \hline \square \\ \hline \square \\ \hline \square \\ \hline \end{array} & \text{1-dimensional antisymmetric representation.}
 \end{array} \tag{B.1.2}$$

The representations of the group S_N are given by the Young tableaux with N boxes. The boxes are arranged in left-bound rows, such that no row is longer than the one above it. For example, for the representations of S_4 one finds


(B.1.3)

The dimension of a representation is determined as follows. The boxes of the corresponding Young tableau are enumerated from 1 to N such that the numbers grow as one reads each row from left to right, and each column from top to bottom. The number of possible enumerations determines the dimension of the representation. For example, for S_3 one obtains

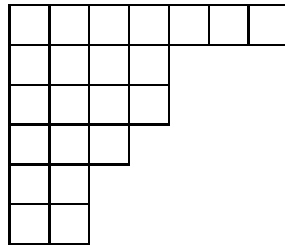

(B.1.4)

The squares of the dimensions of all representations add up to the order of the group, i.e.

$$\sum_{\Gamma} d_{\Gamma}^2 = N! . \quad (B.1.5)$$

In particular, for S_2 we have $1^2 + 1^2 = 2 = 2!$ and for S_3 one obtains $1^2 + 2^2 + 1^2 = 6 = 3!$.

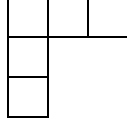
A general Young tableau can be characterized by the number of boxes m_i in its i -th row. For example the Young tableau



has $m_1 = 7$, $m_2 = 4$, $m_3 = 4$, $m_4 = 3$, $m_5 = 2$ and $m_6 = 2$. The dimension of the corresponding representation is given by

$$d_{m_1, m_2, \dots, m_n} = N! \frac{\prod_{i < k} (l_i - l_k)}{l_1! l_2! \dots l_n!}, \quad l_i = m_i + n - i. \quad (\text{B.1.6})$$

Applying this formula to the following Young tableau from S_5



with $m_1 = 3$, $m_2 = 1$, $m_3 = 1$ and $n = 3$ yields $l_1 = 3+3-1 = 5$, $l_2 = 1+3-2 = 2$, $l_3 = 1+3-3 = 1$ and hence

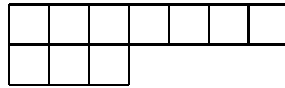
$$d_{3,1,1} = 5! \frac{(l_1 - l_2)(l_1 - l_3)(l_2 - l_3)}{l_1! l_2! l_3!} = 5! \frac{3 \cdot 4 \cdot 1}{5! 2! 1!} = 6. \quad (\text{B.1.7})$$

It is a good exercise to construct all representations of S_4 and S_5 , and determine their dimensions by the enumeration method of eq.(B.1.4) and independently from eq.(B.1.6). One can check the results using eq.(B.1.5).

The permuted objects can be the fundamental representations of $SU(n)$. For $SU(2)$ we identify

$$\square = \{2\}. \quad (\text{B.1.8})$$

To each Young tableau with no more than two rows one can associate an $SU(2)$ representation. Such a Young tableau is characterized by m_1 and m_2 , e.g.



has $m_1 = 7$ and $m_2 = 3$. The corresponding $SU(2)$ representation has

$$S = \frac{1}{2}(m_1 - m_2), \quad (\text{B.1.9})$$

which is also denoted by $\{m_1 - m_2 + 1\}$. The above Young tableau hence represents $S = 2$ — a spin quintet $\{5\}$. Young tableaux with more than two rows have no realization in $SU(2)$ since among just two distinguishable objects no more than two can be combined anti-symmetrically.

B.2 The Group $SU(n)$

The unitary $n \times n$ matrices with determinant 1 form a group under matrix multiplication — the special unitary group $SU(n)$. This follows immediately from

$$\begin{aligned} UU^\dagger &= U^\dagger U = 1, \det U = 1. \\ \det UV &= \det U \det V = 1. \end{aligned} \quad (\text{B.2.1})$$

Associativity ($(UV)W = U(VW)$) holds for all matrices, a unit element 1 exists (the unit matrix), the inverse is $U^{-1} = U^\dagger$, and finally the group property

$$(UV)^\dagger UV = V^\dagger U^\dagger UV = 1, \quad UV(UV)^\dagger = UVV^\dagger U^\dagger = 1 \quad (\text{B.2.2})$$

also holds. The group $SU(n)$ is non-Abelian because in general $UV \neq VU$. Each element $U \in SU(n)$ can be represented as

$$U = \exp(iH), \quad (\text{B.2.3})$$

where H is Hermitean and traceless. The matrices H form the $su(n)$ algebra. One has $n^2 - 1$ free parameters, and hence $n^2 - 1$ generators η_i , and one can write

$$H = \alpha_i \eta_i, \quad \alpha_i \in \mathbb{R}. \quad (\text{B.2.4})$$

The structure of the algebra results from the commutation relations

$$[\eta_i, \eta_j] = 2i c_{ijk} \eta_k, \quad (\text{B.2.5})$$

where c_{ijk} are the so-called structure constants.

The simplest nontrivial representation of $SU(n)$ is the fundamental representation. It is n -dimensional and can be identified with the Young tableau \square . Every irreducible representation of $SU(n)$ can be obtained from coupling N fundamental representations. In this way each $SU(n)$ representation is associated with a Young tableau with N boxes, which characterizes the permutation symmetry of the fundamental representations in the coupling. Since the fundamental representation is n -dimensional, there are n different fundamental properties (e.g. u and d in $SU(2)_I$ and r, g and b in $SU(3)_C$). Hence, we can maximally antisymmetrize n objects, and the Young tableaux for $SU(n)$ representations are therefore restricted to no more than n rows.

The dimension of an $SU(n)$ representation can be obtained from the corresponding Young tableau by filling it with factors as follows

n	$n+1$	$n+2$	$n+3$	$n+4$	$n+5$	$n+6$
$n-1$	n	$n+1$	$n+2$			
$n-2$	$n-1$	n	$n+1$			
$n-3$	$n-2$	$n-1$				
$n-4$	$n-3$					
$n-5$	$n-4$					

The dimension of the $SU(n)$ representation is given as the product of all factors divided by $N!$ and multiplied with the S_N dimension d_{m_1, m_2, \dots, m_n} of the Young tableau

$$\begin{aligned}
 D_{m_1, m_2, \dots, m_n}^n &= \frac{(n+m_1-1)!}{(n-1)!} \frac{(n+m_2-2)!}{(n-2)!} \dots \frac{m_n!}{0!} \frac{1}{N!} N! \frac{\prod_{i < k} (l_i - l_k)}{l_1! l_2! \dots l_n!} \\
 &= \frac{\prod_{i < k} (m_i - m_k - i + k)}{(n-1)!(n-2)! \dots 0!}.
 \end{aligned} \tag{B.2.6}$$

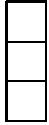
We see that the dimension of a representation depends only on the differences $q_i = m_i - m_{i+1}$. In particular, for $SU(2)$ we find

$$D_{m_1, m_2}^2 = \frac{m_1 - m_2 - 1 + 2}{1!0!} = m_1 - m_2 + 1 = q_1 + 1 \tag{B.2.7}$$

in agreement with our previous result. For a rectangular Young tableau with n rows, e.g. in $SU(2)$ for

all $q_i = 0$, and we obtain

$$D_{m, m, \dots, m}^n = \frac{\prod_{i < k} (m_i - m_k - i + k)}{(n-1)!(n-2)! \dots 0!} = \frac{(n-1)!(n-2)! \dots 0!}{(n-1)!(n-2)! \dots 0!} = 1, \tag{B.2.8}$$

and therefore a singlet. This confirms that in $SU(3)$  corresponds to a singlet. It also explains why the dimension of an $SU(n)$ representation depends only on the differences q_i . Without changing the dimension we can couple a representation with a singlet, and hence we can always add a rectangular Young tableau with n rows to any $SU(n)$ representation. For example in $SU(3)$

$$\begin{array}{|c|c|c|} \hline & & \\ \hline & & \\ \hline & & \\ \hline \end{array} \cong \begin{array}{|c|c|} \hline & \\ \hline & \\ \hline \end{array} \quad (\text{B.2.9})$$

We want to associate an anti-representation with each representation by replacing m_i and q_i with

$$\bar{m}_i = m_1 - m_{n-i+1}, \quad \bar{q}_i = \bar{m}_i - \bar{m}_{i+1} = m_{n-i} - m_{n-i+1} = q_{n-i}. \quad (\text{B.2.10})$$

Geometrically the Young tableau of a representation and its anti-representation (after rotation) fit together to form a rectangular Young tableau with n rows. For example, in $SU(3)$

$$\begin{array}{|c|c|c|} \hline & & \\ \hline & & \\ \hline & & \\ \hline \end{array} \quad \text{and} \quad \begin{array}{|c|c|} \hline & \\ \hline & \\ \hline & \\ \hline \end{array}$$

are anti-representations of one another. In $SU(2)$ each representation is its own anti-representation. For example

$$\begin{array}{|c|c|c|} \hline & & \\ \hline & & \\ \hline & & \\ \hline \end{array} \quad \text{and} \quad \begin{array}{|c|c|} \hline & \\ \hline & \\ \hline \end{array}$$

are anti-representations of one another, but

$$\begin{array}{|c|c|c|} \hline & & \\ \hline & & \\ \hline & & \\ \hline \end{array} \cong \begin{array}{|c|c|} \hline & \\ \hline & \\ \hline \end{array}.$$

This is not the case for higher n . The dimension of a representation and its anti-representation are identical

$$D_{\bar{m}_1, \bar{m}_2, \dots, \bar{m}_n}^n = D_{m_1, m_2, \dots, m_n}^n. \quad (\text{B.2.11})$$

It is an instructive exercise to prove this. For general n the so-called adjoint representation is given by $q_1 = q_{n-1} = 1$, $q_i = 0$ otherwise, and it is identical with

its own anti-representation. Again it is instructive to show that the dimension of the adjoint representation is

$$D_{2,1,1,\dots,1,0}^n = n^2 - 1. \quad (\text{B.2.12})$$

Next we want to discuss a method to couple $SU(n)$ representations by operating on their Young tableaux. Two Young tableaux with N and M boxes are coupled by forming an external product. In this way we generate Young tableaux with $N + M$ boxes that can then be translated back into $SU(n)$ representations. The external product is built as follows. The boxes of the first row of the second Young tableau are labeled with 'a', the boxes of the second row with 'b', etc. Then the boxes labeled with 'a' are added to the first Young tableau in all possible ways that lead to new allowed Young tableau. Then the 'b' boxes are added to the resulting Young tableaux in the same way. Now each of the resulting tableaux is read row-wise from top-right to bottom-left. Whenever a 'b' or 'c' appears before the first 'a', or a 'c' occurs before the first 'b' etc., the corresponding Young tableau is deleted. The remaining tableaux form the reduction of the external product. It is instructive to couple

$$\begin{array}{|c|c|} \hline \square & \square \\ \hline \square & \square \\ \hline \end{array} \otimes \begin{array}{|c|c|} \hline \square & \square \\ \hline \square & \square \\ \hline \end{array} \quad (\text{B.2.13})$$

in this way. As a check one can translate the result into $SU(2)$ and $SU(3)$ language and confirm that the product of the dimensions of the coupled representations agrees with the sum of dimensions of the representations in the reduction.

We now want to couple N fundamental representations of $SU(n)$. In Young tableau language this reads

$$\{n\} \otimes \{n\} \otimes \dots \otimes \{n\} = \square \otimes \square \otimes \dots \otimes \square. \quad (\text{B.2.14})$$

In this way we generate all irreducible representations of S_N , i.e. all Young tableaux with N boxes. Each Young tableau is associated with an $SU(n)$ multiplet. It occurs in the product as often as the dimension of the corresponding S_N representation indicates, i.e. d_{m_1, m_2, \dots, m_n} times. Hence we can write

$$\{n\} \otimes \{n\} \otimes \dots \otimes \{n\} = \sum_{\Gamma} d_{m_1, m_2, \dots, m_n} \{D_{m_1, m_2, \dots, m_n}^n\}. \quad (\text{B.2.15})$$

The sum goes over all Young tableaux with N boxes. For example

$$\square \otimes \square \otimes \square = \begin{array}{|c|c|c|} \hline \square & \square & \square \\ \hline \end{array} \oplus 2 \begin{array}{|c|c|} \hline \square & \square \\ \hline \square & \square \\ \hline \end{array} \oplus \begin{array}{|c|} \hline \square \\ \hline \square \\ \hline \square \\ \hline \end{array}. \quad (\text{B.2.16})$$

Translated into $SU(n)$ language this reads

$$\begin{aligned} \{n\} \otimes \{n\} \otimes \{n\} &= \left\{ \frac{n(n+1)(n+2)}{6} \right\} \oplus 2 \left\{ \frac{(n-1)n(n+1)}{3} \right\} \\ &\oplus \left\{ \frac{(n-2)(n-1)n}{6} \right\}. \end{aligned} \quad (\text{B.2.17})$$

The dimensions test

$$\frac{n(n+1)(n+2)}{6} + 2 \frac{(n-1)n(n+1)}{3} + \frac{(n-2)(n-1)n}{6} = n^3 \quad (\text{B.2.18})$$

confirms this result. In $SU(2)$ this corresponds to

$$\{2\} \otimes \{2\} \otimes \{2\} = \{4\} \oplus 2\{2\} \oplus \{0\}, \quad (\text{B.2.19})$$

and in $SU(3)$

$$\{3\} \otimes \{3\} \otimes \{3\} = \{10\} \oplus 2\{8\} \oplus \{1\}. \quad (\text{B.2.20})$$

As an exercise using the above technique one can couple the fundamental representation with its anti-representation, and show that one obtains a singlet and the adjoint representation for any $SU(n)$, i.e.

$$\{\bar{n}\} \otimes \{n\} = \{1\} \oplus \{n^2 - 1\}. \quad (\text{B.2.21})$$

In $SU(2)$ this reads $\{2\} \otimes \{2\} = \{1\} \oplus \{3\}$ and in $SU(3)$ we have $\{\bar{3}\} \otimes \{3\} = \{1\} \oplus \{8\}$.

Appendix C

Topology of Gauge Fields

C.1 The Anomaly

Let us consider the baryon number current in the standard model

$$J_\mu(x) = \bar{q}(x)\gamma_\mu q(x). \quad (\text{C.1.1})$$

The Lagrange density of the standard model is invariant under global $U(1)$ baryon number transformations. The corresponding Noether current J_μ is hence conserved at the classical level

$$\partial^\mu J_\mu(x) = 0. \quad (\text{C.1.2})$$

At the quantum level, however, the symmetry cannot be maintained because it is violated by the Adler-Bell-Jackiw anomaly

$$\partial^\mu J_\mu(x) = N_g P(x). \quad (\text{C.1.3})$$

Here N_g is the number of generations, and

$$P(x) = -\frac{1}{32\pi^2} \varepsilon^{\mu\nu\rho\sigma} \text{Tr}(G_{\mu\nu}(x)G_{\rho\sigma}(x)) \quad (\text{C.1.4})$$

is the so-called Chern-Pontryagin density. In the standard model $G_{\mu\nu}$ is the field strength tensor of the W -bosons. In the following we consider the topology of a general non-Abelian gauge potential G_μ . The anomaly equation can be derived in perturbation theory and it follows from a triangle diagram. The Chern-Pontryagin density can be written as a total divergence

$$P(x) = \partial^\mu \Omega_\mu^{(0)}(x), \quad (\text{C.1.5})$$

where $\Omega_\mu^{(0)}(x)$ is the so-called Chern-Simons density or 0-cochain, which is given by

$$\Omega_\mu^{(0)}(x) = -\frac{1}{8\pi^2} \varepsilon^{\mu\nu\rho\sigma} \text{Tr}[G_\nu(x)(\partial_\rho G_\sigma(x) + \frac{2}{3} G_\rho(x)G_\sigma(x))]. \quad (\text{C.1.6})$$

It is a good exercise to convince oneself that this satisfies eq.(C.1.5). We can now formally construct a conserved current

$$\tilde{J}_\mu(x) = J_\mu(x) - N_g \Omega_\mu^{(0)}(x), \quad (\text{C.1.7})$$

because then

$$\partial^\mu \tilde{J}_\mu(x) = \partial^\mu J_\mu(x) - N_g P(x) = 0. \quad (\text{C.1.8})$$

One might think that we have found a new $U(1)$ symmetry which is free of the anomaly. This is, however, not the case, because the current $\tilde{J}_\mu(x)$ contains $\Omega_\mu^{(0)}(x)$ which is not gauge invariant. Although the gauge variant current is formally conserved, this has no gauge invariant physical consequences.

C.2 The Topological Charge

For the rest of this chapter we will leave Minkowski space-time and Wick rotate ourselves into a Euclidean world with an imaginary (or Euclidean) time. It is then difficult to interpret space-time processes, because we have to perform an analytic continuation to make contact with the real world. Still, it is mathematically advantageous to use Euclidean time, and physical results like particle masses remain unaffected when we go back to Minkowski space-time. From now on we stop distinguishing co- and contravariant indices and we write the topological charge as

$$\begin{aligned} Q &= -\frac{1}{32\pi^2} \int d^4x \varepsilon_{\mu\nu\rho\sigma} \text{Tr}(G_{\mu\nu}(x)G_{\rho\sigma}(x)) = \int d^4x P(x) \\ &= \int d^4x \partial_\mu \Omega_\mu^{(0)}(x) = \int_{S^3} d^3\sigma_\mu \Omega_\mu^{(0)}(x). \end{aligned} \quad (\text{C.2.1})$$

We have used Gauss' law to reduce the integral over Euclidean space-time to an integral over its boundary, which is topologically a 3-sphere S^3 . We will restrict ourselves to gauge field configurations with a finite action. Hence, their field strength should vanish at infinity, and consequently the gauge potential should then be a pure gauge (a gauge transformation of a zero field)

$$G_\mu(x) = g(x)\partial_\mu g(x)^\dagger. \quad (\text{C.2.2})$$

Of course, this expression is only valid at space-time infinity. Inserting it in the expression for the 0-cochain we obtain

$$\begin{aligned}
Q &= -\frac{1}{8\pi^2} \int_{S^3} d^3\sigma_\mu \varepsilon_{\mu\nu\rho\sigma} \text{Tr}[(g(x)\partial_\nu g(x)^\dagger)(\partial_\rho(g(x)\partial_\sigma g(x)^\dagger) \\
&\quad + \frac{2}{3}(g(x)\partial_\rho g(x)^\dagger)(g(x)\partial_\sigma g(x)^\dagger))] \\
&= -\frac{1}{8\pi^2} \int_{S^3} d^3\sigma_\mu \varepsilon_{\mu\nu\rho\sigma} \\
&\quad \times \text{Tr}[-(g(x)\partial_\nu g(x)^\dagger)(g(x)\partial_\rho g(x)^\dagger)(g(x)\partial_\sigma g(x)^\dagger) \\
&\quad + \frac{2}{3}(g(x)\partial_\nu g(x)^\dagger)(g(x)\partial_\rho g(x)^\dagger)(g(x)\partial_\sigma g(x)^\dagger)] \\
&= \frac{1}{24\pi^2} \int_{S^3} d^3\sigma_\mu \varepsilon_{\mu\nu\rho\sigma} \text{Tr}[(g(x)\partial_\nu g(x)^\dagger)(g(x)\partial_\rho g(x)^\dagger)(g(x)\partial_\sigma g(x)^\dagger)].
\end{aligned} \tag{C.2.3}$$

The gauge transformation $g(x)$ defines a mapping of the sphere S^3 at space-time infinity to the gauge group $SU(N)$

$$g : S^3 \rightarrow SU(N). \tag{C.2.4}$$

Such mappings have topological properties. They fall into equivalence classes — the so-called homotopy classes — which represent topologically distinct sectors. Two mappings are equivalent if they can be deformed continuously into one another. The homotopy properties are described by so-called homotopy groups. In our case the relevant homotopy group is

$$\Pi_3[SU(N)] = \mathbb{Z}. \tag{C.2.5}$$

Here the index 3 indicates that we consider mappings of the 3-dimensional sphere S^3 . The third homotopy group of $SU(N)$ is given by the integers. This means that for each integer Q there is a class of mappings that can be continuously deformed into one another, while mappings with different Q are topologically distinct. The integer Q that characterizes the mapping topologically is the topological charge. Now we want to show that the above expression for Q is exactly that integer. For this purpose we decompose

$$g = VW, \quad W = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & \tilde{g}_{11} & \tilde{g}_{12} & \dots & \tilde{g}_{1N} \\ 0 & \tilde{g}_{21} & \tilde{g}_{22} & \dots & \tilde{g}_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \tilde{g}_{N1} & \tilde{g}_{N2} & \dots & \tilde{g}_{NN} \end{pmatrix}, \tag{C.2.6}$$

where the embedded matrix \tilde{g} is in $SU(N-1)$. It is indirectly defined by

$$V = \begin{pmatrix} g_{11} & -g_{21}^* & -\frac{g_{31}^*(1+g_{11})}{1+g_{11}^*} & \cdots & -\frac{g_{N1}^*(1+g_{11})}{1+g_{11}^*} \\ g_{21} & \frac{1+g_{11}^*|g_{21}|^2}{1+g_{11}} & -\frac{g_{31}^*g_{21}}{1+g_{11}^*} & \cdots & -\frac{g_{N1}^*g_{21}}{1+g_{11}^*} \\ g_{31} & -\frac{g_{21}^*g_{31}}{1+g_{11}} & \frac{1+g_{11}^*|g_{31}|^2}{1+g_{11}^*} & \cdots & -\frac{g_{N1}^*g_{31}}{1+g_{11}^*} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_{N1} & -\frac{g_{21}^*g_{N1}}{1+g_{11}} & -\frac{g_{31}^*g_{N1}}{1+g_{11}^*} & \cdots & \frac{1+g_{11}^*|g_{N1}|^2}{1+g_{11}^*} \end{pmatrix} \in SU(N). \quad (\text{C.2.7})$$

The matrix V is constructed entirely from the elements $g_{11}, g_{21}, \dots, g_{N1}$ of the first column of the matrix g . One should convince oneself that V is indeed an $SU(N)$ matrix, and that the resulting matrix \tilde{g} is indeed in $SU(N-1)$. The idea now is to reduce the expression for the topological charge from $SU(N)$ to $SU(N-1)$ by using the formula

$$\begin{aligned} & \varepsilon_{\mu\nu\rho\sigma} \text{Tr}[(VW)\partial_\nu(VW)^\dagger(VW)\partial_\rho(VW)^\dagger(VW)\partial_\sigma(VW)^\dagger] = \\ & \varepsilon_{\mu\nu\rho\sigma} \text{Tr}[(V\partial_\nu V^\dagger)(V\partial_\rho V^\dagger)(V\partial_\sigma V^\dagger) \\ & + \varepsilon_{\mu\nu\rho\sigma} \text{Tr}[(W\partial_\nu W^\dagger)(W\partial_\rho W^\dagger)(W\partial_\sigma W^\dagger)] \\ & + 3\partial_\nu \varepsilon_{\mu\nu\rho\sigma} \text{Tr}[(V\partial_\rho V^\dagger)(W\partial_\sigma W^\dagger)]. \end{aligned} \quad (\text{C.2.8})$$

Again, it is instructive to prove this formula. Applying the formula to the expression for the topological charge and using $g = VW$ we obtain

$$\begin{aligned} Q &= \frac{1}{24\pi^2} \int_{S^3} d^3\sigma_\mu \varepsilon_{\mu\nu\rho\sigma} \text{Tr}[(g(x)\partial_\nu g(x)^\dagger)(g(x)\partial_\rho g(x)^\dagger)(g(x)\partial_\sigma g(x)^\dagger)] \\ &= \frac{1}{24\pi^2} \int_{S^3} d^3\sigma_\mu \varepsilon_{\mu\nu\rho\sigma} \text{Tr}[(V(x)\partial_\nu V(x)^\dagger)(V(x)\partial_\rho V(x)^\dagger)(V(x)\partial_\sigma V(x)^\dagger) \\ &+ (W(x)\partial_\nu W(x)^\dagger)(W(x)\partial_\rho W(x)^\dagger)(W(x)\partial_\sigma W(x)^\dagger)]. \end{aligned} \quad (\text{C.2.9})$$

The ∂_ν term of the formula eq.(C.2.8) drops out using Gauss' law together with the fact that S^3 has no boundary. It follows that the topological charge of a product of two gauge transformations V and W is the sum of the topological charges of V and W . Since V only depends on $g_{11}, g_{21}, \dots, g_{N1}$, it can be viewed as a mapping of S^3 into the sphere S^{2N-1}

$$V : S^3 \rightarrow S^{2N-1}. \quad (\text{C.2.10})$$

This is because $|g_{11}|^2 + |g_{21}|^2 + \dots + |g_{N1}|^2 = 1$. Remarkably the corresponding homotopy group is trivial for $N > 2$, i.e.

$$\Pi_3[S^{2N-1}] = \{0\}. \quad (\text{C.2.11})$$

All mappings of S^3 into the higher dimensional sphere S^{2N-1} are topologically equivalent (they can be deformed into each other). This can be understood better in a lower dimensional example

$$\Pi_1[S^2] = \{0\}. \quad (\text{C.2.12})$$

Each closed curve on an ordinary sphere can be constricted to the north pole, and hence is topologically trivial. In fact,

$$\Pi_m[S^n] = \{0\}, \quad (\text{C.2.13})$$

for $m < n$, while

$$\Pi_n[S^n] = \mathbb{Z}. \quad (\text{C.2.14})$$

Still, $\Pi_m[S^n]$ with $m > n$ is not necessarily trivial, for example

$$\Pi_4[S^3] = \mathbb{Z}(2). \quad (\text{C.2.15})$$

Since the mapping V is topologically trivial its contribution to the topological charge vanishes. The remaining W term reduces to the $SU(N-1)$ contribution

$$Q = \frac{1}{24\pi^2} \int_{S^3} d^3\sigma_\mu \varepsilon_{\mu\nu\rho\sigma} \text{Tr}[(\tilde{g}(x)\partial_\nu\tilde{g}(x)^\dagger)(\tilde{g}(x)\partial_\rho\tilde{g}(x)^\dagger)(\tilde{g}(x)\partial_\sigma\tilde{g}(x)^\dagger)]. \quad (\text{C.2.16})$$

The separation of the V contribution works only if the decomposition of g into V and \tilde{g} is non-singular. In fact, the expression for V is singular for $g_{11} = -1$. This corresponds to a $((N-1)^2 - 1)$ -dimensional subspace of the $(N^2 - 1)$ -dimensional $SU(N)$ group space. The mapping g itself covers a 3-d subspace of $SU(N)$. Hence it is arbitrarily improbable to hit a singularity (it is of zero measure). Since we have now reduced the $SU(N)$ topological charge to the $SU(N-1)$ case, we can go down all the way to $SU(2)$. It remains to be shown that the $SU(2)$ expression is actually an integer. First of all

$$\tilde{g} : S^3 \rightarrow SU(2) = S^3, \quad (\text{C.2.17})$$

and indeed

$$\Pi_3[SU(2)] = \Pi_3[S^3] = \mathbb{Z}. \quad (\text{C.2.18})$$

The topological charge specifies how often the $SU(2)$ group space (which is isomorphic to the 3-sphere) is covered by \tilde{g} as we go along the boundary of Euclidean space-time (which is also topologically S^3). Again, it is useful to consider a lower dimensional example, mappings from the circle S^1 to the group $U(1)$ which is topologically also a circle

$$g = \exp(i\varphi) : S^1 \rightarrow U(1) = S^1. \quad (\text{C.2.19})$$

The relevant homotopy group is

$$\Pi_1[U(1)] = \Pi_1[S^1] = \mathbb{Z}. \quad (\text{C.2.20})$$

Again, for each integer there is an equivalence class of mappings that can be continuously deformed into one another. Going over the circle S^1 the mapping may cover the group space $U(1)$ any number of times. In $U(1)$ the expression for the topological charge is analogous to the one in $SU(N)$

$$\begin{aligned} Q &= -\frac{1}{2\pi} \int_{S^1} d\sigma_\mu \varepsilon_{\mu\nu} (g(x) \partial_\nu g(x)^\dagger) = \frac{1}{2\pi} \int_{S^1} d\sigma_\mu \varepsilon_{\mu\nu} \partial_\nu \varphi(x) \\ &= \frac{1}{2\pi} (\varphi(2\pi) - \varphi(0)). \end{aligned} \quad (\text{C.2.21})$$

If $g(x)$ is continuous over the circle $\varphi(2\pi)$ and $\varphi(0)$ must differ by 2π times an integer. That integer is the topological charge. It counts how many times the mapping g covers the group space $U(1)$ as we move along the circle S^1 . We are looking for an analogous expression in $SU(2)$. For this purpose we parametrize the mapping \tilde{g} as

$$\begin{aligned} \tilde{g}(x) &= \exp(i\vec{\alpha}(x) \cdot \vec{\sigma}) = \cos \alpha(x) + i \sin \alpha(x) \vec{e}_\alpha(x) \cdot \vec{\sigma}, \\ \vec{e}_\alpha(x) &= (\sin \theta(x) \sin \varphi(x), \sin \theta(x) \cos \varphi(x), \cos \theta(x)). \end{aligned} \quad (\text{C.2.22})$$

It is a good exercise to convince oneself that

$$\begin{aligned} &\varepsilon_{\mu\nu\rho\sigma} \text{Tr}[(\tilde{g}(x) \partial_\nu \tilde{g}(x)^\dagger)(\tilde{g}(x) \partial_\rho \tilde{g}(x)^\dagger)(\tilde{g}(x) \partial_\sigma \tilde{g}(x)^\dagger)] \\ &= 12 \sin^2 \alpha(x) \sin \theta(x) \varepsilon_{\mu\nu\rho\sigma} \partial_\nu \alpha(x) \partial_\rho \theta(x) \partial_\sigma \varphi(x). \end{aligned} \quad (\text{C.2.23})$$

This is exactly the volume element of a 3-sphere (and hence of the $SU(2)$ group space). Thus we can now write

$$Q = \frac{1}{2\pi^2} \int_{S^3} d^3\sigma_\mu \sin^2 \alpha(x) \sin \theta(x) \varepsilon_{\mu\nu\rho\sigma} \partial_\nu \alpha(x) \partial_\rho \theta(x) \partial_\sigma \varphi(x) = \frac{1}{2\pi^2} \int_{S^3} d\tilde{g}. \quad (\text{C.2.24})$$

The volume of the 3-sphere is given by $2\pi^2$. When the mapping \tilde{g} covers the sphere Q times the integral gives Q times the volume of S^3 . This finally explains why the prefactor $1/32\pi^2$ was introduced in the original expression for the topological charge.

C.3 Gauge Field Topology on a Compact Manifold

Imagine our Universe was closed both in space and time, and hence had no boundary. Our previous discussion, for which the value of the gauge field at the

boundary was essential, would suggest that in a closed Universe the topology is trivial. On the other hand, we think that topology has local consequences. For example, baryon number is violated because the topological charge does not vanish. To resolve this apparent contradiction we will now discuss the topology of a gauge field on a compact Euclidean space-time manifold M , and we will see that nontrivial topology is still present. Let us again consider the topological charge

$$Q = \int_M d^4x P(x). \quad (\text{C.3.1})$$

Writing the Chern-Pontryagin density as the total divergence of the 0-cochain

$$P(x) = \partial_\mu \Omega_\mu^{(0)}(x), \quad (\text{C.3.2})$$

and using Gauss' law we obtain

$$Q = \int_M d^4x \partial_\mu \Omega_\mu^{(0)}(x) = \int_{\partial M} d^3\sigma_\mu \Omega_\mu^{(0)}(x) = 0. \quad (\text{C.3.3})$$

Here we have used that M has no boundary, i.e. ∂M is an empty set. A gauge field whose Chern-Pontryagin density can globally be written as a total divergence is indeed topologically trivial on a compact manifold. The important observation is that eq.(C.3.2) may be valid only locally. In other words, gauge singularities may prevent us from using Gauss' law as we did above. In general, it will be impossible to work in a gauge that makes the gauge field nonsingular everywhere on the space-time manifold. Instead we must subdivide space-time into local patches in which the gauge field is smooth, and glue the patches together by nontrivial gauge transformations, which form a fibre bundle of transition functions. A topologically nontrivial gauge field will contain singularities at some points $x_i \in M$. We cover the manifold M by closed sets c_i such that $x_i \in c_i \setminus \partial c_i$, i.e. each singularity lies in the interior of a set c_i . Also $M = \cup_i c_i$ with $c_i \cap c_j = \partial c_i \cap \partial c_j$.

The next step is to remove the gauge singularities x_i by performing gauge transformations g_i in each local patch

$$G_\mu^i(x) = g_i(x)(G_\mu(x) + \partial_\mu)g_i^\dagger(x). \quad (\text{C.3.4})$$

After the gauge transformation the gauge potential $G_\mu^i(x)$ is free of singularities in the local region c_i . Hence we can now use Gauss' law and obtain

$$\begin{aligned} Q &= \sum_i \int_{c_i} d^4x P(x) = \sum_i \int_{\partial c_i} d^3\sigma_\mu \Omega_\mu^{(0)}(i) \\ &= \frac{1}{2} \sum_{ij} \int_{c_i \cap c_j} d^3\sigma_\mu [\Omega_\mu^{(0)}(i) - \Omega_\mu^{(0)}(j)]. \end{aligned} \quad (\text{C.3.5})$$

The argument i of the 0-cochain indicates that we are in the region c_i . At the intersection of two regions $c_i \cap c_j$ the gauge field G_μ^i differs from G_μ^j , although the original gauge field $G_\mu(x)$ was continuous there. In fact, the two gauge fields are related by a gauge transformation v_{ij}

$$G_\mu^i(x) = v_{ij}(x)(G_\mu^j(x) + \partial_\mu)v_{ij}(x)^\dagger, \quad (\text{C.3.6})$$

which is defined only on $c_i \cap c_j$. The gauge transformations v_{ij} form a fibre bundle of transition functions given by

$$v_{ij}(x) = g_i(x)g_j(x)^\dagger. \quad (\text{C.3.7})$$

This equation immediately implies a consistency equation. This so-called cocycle condition relates the transition functions in the intersection $c_i \cap c_j \cap c_k$ of three regions

$$v_{ik}(x) = v_{ij}(x)v_{jk}(x). \quad (\text{C.3.8})$$

The above difference of two 0-cochains in different gauges is given by the so-called coboundary operator Δ

$$\Delta\Omega_\mu^{(0)}(i, j) = \Omega_\mu^{(0)}(i) - \Omega_\mu^{(0)}(j). \quad (\text{C.3.9})$$

It is straight forward to show that

$$\begin{aligned} \Delta\Omega_\mu^{(0)}(i, j) &= -\frac{1}{24\pi^2}\varepsilon_{\mu\nu\rho\sigma}\text{Tr}[v_{ij}(x)\partial_\nu v_{ij}(x)^\dagger v_{ij}(x)\partial_\rho v_{ij}(x)^\dagger v_{ij}(x)\partial_\sigma v_{ij}(x)^\dagger] \\ &\quad -\frac{1}{8\pi^2}\varepsilon_{\mu\nu\rho\sigma}\partial_\nu\text{Tr}[\partial_\rho v_{ij}(x)^\dagger v_{ij}(x)G_\sigma^i(x)]. \end{aligned} \quad (\text{C.3.10})$$

The above equation for the topological charge then takes the form

$$\begin{aligned} Q &= -\frac{1}{48\pi^2}\sum_{ij}\int_{c_i\cap c_j}d^3\sigma_\mu\varepsilon_{\mu\nu\rho\sigma} \\ &\quad \times \text{Tr}[v_{ij}(x)\partial_\nu v_{ij}(x)^\dagger v_{ij}(x)\partial_\rho v_{ij}(x)^\dagger v_{ij}(x)\partial_\sigma v_{ij}(x)^\dagger] \\ &\quad -\frac{1}{16\pi^2}\sum_{ij}\int_{\partial(c_i\cap c_j)}d^2\sigma_{\mu\nu}\varepsilon_{\mu\nu\rho\sigma}\text{Tr}[\partial_\rho v_{ij}(x)^\dagger v_{ij}(x)G_\sigma^i(x)]. \end{aligned} \quad (\text{C.3.11})$$

Using the cocycle condition this can be rewritten as

$$\begin{aligned} Q &= -\frac{1}{48\pi^2}\sum_{ij}\int_{c_i\cap c_j}d^3\sigma_\mu\varepsilon_{\mu\nu\rho\sigma} \\ &\quad \times \text{Tr}[v_{ij}(x)\partial_\nu v_{ij}(x)^\dagger v_{ij}(x)\partial_\rho v_{ij}(x)^\dagger v_{ij}(x)\partial_\sigma v_{ij}(x)^\dagger] \\ &\quad -\frac{1}{48\pi^2}\sum_{ijk}\int_{c_i\cap c_j\cap c_k}d^2\sigma_{\mu\nu}\varepsilon_{\mu\nu\rho\sigma}\text{Tr}[v_{ij}(x)\partial_\rho v_{ij}(x)^\dagger v_{jk}(x)\partial_\rho v_{jk}(x)^\dagger]. \end{aligned} \quad (\text{C.3.12})$$

This shows that the topology of the fibre bundle is entirely encoded in the transition functions.

In the appropriate mathematical language the gauge transformations g_i form sections of the fibre bundle. Using formula (C.2.8) together with eq.(C.3.7) one can show that the topological charge is expressed in terms of the section in the following way

$$\begin{aligned} Q &= \sum_i Q_i \\ &= \frac{1}{24\pi^2} \sum_i \int_{\partial c_i} d^3 \sigma_\mu \varepsilon_{\mu\nu\rho\sigma} \text{Tr}[g_i(x) \partial_\nu g_i(x)^\dagger g_i(x) \partial_\rho g_i(x)^\dagger g_i(x) \partial_\sigma g_i(x)^\dagger]. \end{aligned} \quad (\text{C.3.13})$$

We recognize the integer winding number Q_i that characterizes the mapping g_i topologically. In fact, the boundary ∂c_i is topologically a 3-sphere, such that

$$g_i : \partial c_i \rightarrow SU(3), \quad (\text{C.3.14})$$

and hence

$$Q_i \in \Pi_3[SU(3)] = \mathbb{Z}. \quad (\text{C.3.15})$$

The topological charge Q is a sum of local winding numbers $Q_i \in \mathbb{Z}$, which are associated with the regions c_i . In general, the Q_i are not gauge invariant. Hence, individually they have no physical meaning. Still, the total charge — as the sum of all Q_i — is gauge invariant. It is instructive to show this explicitly by performing a gauge transformation on the original gauge field

$$G_\mu(x)' = g(x)(G_\mu(x) + \partial_\mu)g(x)^\dagger. \quad (\text{C.3.16})$$

Deriving the gauge transformation properties of the section and using formula (C.2.8) this is again straightforward.

C.4 Cochain Reduction in $SU(2)$

To complete the mathematical investigation of the topological charge let us finally investigate its expression eq.(C.3.12) in terms of transition functions. To understand why this expression also is an integer we concentrate on $SU(2)$ and we go back to the equation

$$Q = \sum_i \int_{\partial c_i} d^3 \sigma_\mu \Omega_\mu^{(0)}(i) = \frac{1}{2!} \sum_{ij} \int_{c_i \cap c_j} d^3 \sigma_\mu [\Omega_\mu^{(0)}(i) - \Omega_\mu^{(0)}(j)]. \quad (\text{C.4.1})$$

The coboundary operator

$$\Delta\Omega_\mu^{(0)}(i, j) = \Omega_\mu^{(0)}(i) - \Omega_\mu^{(0)}(j) = \partial_\nu \Omega_{\mu\nu}^{(1)}(i, j) \quad (\text{C.4.2})$$

can be written as a total divergence of the 1-cochain

$$\begin{aligned} \Omega_{\mu\nu}^{(1)}(i, j) &= -\frac{1}{8\pi^2}(\alpha - \sin\alpha \cos\alpha) \varepsilon_{\mu\nu\rho\sigma} \vec{e}_\alpha \cdot (\partial_\rho \vec{e}_\alpha \times \partial_\sigma \vec{e}_\alpha) \\ &\quad - \frac{1}{8\pi^2} \varepsilon_{\mu\nu\rho\sigma} \text{Tr}[\partial_\rho v_{ij} v_{ij}^\dagger G_\sigma^i]. \end{aligned} \quad (\text{C.4.3})$$

Here we have parametrized the $SU(2)$ transition function as

$$v_{ij} = \exp(i\vec{\alpha} \cdot \vec{\sigma}) = \cos\alpha + i \sin\alpha \vec{e}_\alpha \cdot \vec{\sigma}, \quad \alpha \in [0, \pi]. \quad (\text{C.4.4})$$

When $v_{ij} = -1$, i.e. when $\alpha = \pi$, the above parametrization is singular because then the unit vector \vec{e}_α is not well defined. The singularities $v_{ij} = -1$ occur at isolated points $x \in c_i \cap c_j$. In their neighborhood the 1-cochain has a directional singularity

$$\Omega_{\mu\nu}^{(1)}(i, j) = -\frac{1}{8\pi} \varepsilon_{\mu\nu\rho\sigma} \vec{e}_\alpha \cdot (\partial_\rho \vec{e}_\alpha \times \partial_\sigma \vec{e}_\alpha). \quad (\text{C.4.5})$$

Using Gauss' law the topological charge can now be written as

$$Q = Q^{(1)} + Q_\Sigma^{(1)}, \quad (\text{C.4.6})$$

where

$$Q^{(1)} = \frac{1}{2!} \sum_{ij} \sum_{x \in c_i \cap c_j} \frac{1}{8\pi} \int_{S_\varepsilon^2(x)} d^2\sigma_{\mu\nu} \varepsilon_{\mu\nu\rho\sigma} \vec{e}_\alpha \cdot (\partial_\rho \vec{e}_\alpha \times \partial_\sigma \vec{e}_\alpha). \quad (\text{C.4.7})$$

Here $S_\varepsilon^2(x)$ is a 2-sphere of infinitesimal radius ε around the singularity x . It appears as an internal boundary in the application of Gauss' law because the integrand is singular. Performing the 2-d integral gives

$$\frac{1}{8\pi} \int_{S_\varepsilon^2(x)} d^2\sigma_{\mu\nu} \varepsilon_{\mu\nu\rho\sigma} \vec{e}_\alpha \cdot (\partial_\rho \vec{e}_\alpha \times \partial_\sigma \vec{e}_\alpha) = n^{(1)}(x; i, j), \quad (\text{C.4.8})$$

where $n^{(1)}(x; i, j) \in \mathbb{Z}$ is a local winding number associated with the singularity. It is an element of $\Pi_2[S^2] = \mathbb{Z}$ and it counts how often the unit vector \vec{e}_α covers S^2 as we integrate over $S_\varepsilon^2(x)$. Hence

$$Q^{(1)} = \sum_{\Lambda^{(1)}} n^{(1)}(x; i, j) \quad (\text{C.4.9})$$

is a sum of local winding numbers associated with the singular points that form the set

$$\Lambda^{(1)} = \cup_{ij} \{x \in c_i \cap c_j\}. \quad (\text{C.4.10})$$

The remaining contribution to the topological charge is given by

$$\begin{aligned} Q_{\Sigma}^{(1)} &= \frac{1}{2!} \sum_{ij} \int_{\partial(c_i \cap c_j)} d^2 \sigma_{\mu\nu} \Omega_{\mu\nu}^{(1)}(i, j) \\ &= \frac{1}{3!} \sum_{ijk} \int_{c_i \cap c_j \cap c_k} d^2 \sigma_{\mu\nu} [\Omega_{\mu\nu}^{(1)}(i, j) - \Omega_{\mu\nu}^{(1)}(i, k) + \Omega_{\mu\nu}^{(1)}(j, k)]. \end{aligned} \quad (\text{C.4.11})$$

We identify the coboundary operator

$$\Delta \Omega_{\mu\nu}^{(1)}(i, j, k) = \Omega_{\mu\nu}^{(1)}(i, j) - \Omega_{\mu\nu}^{(1)}(i, k) + \Omega_{\mu\nu}^{(1)}(j, k) = \partial_{\rho} \Omega_{\mu\nu\rho}^{(2)}(i, j, k), \quad (\text{C.4.12})$$

which again is a total divergence. The 2-cochain is given by

$$\begin{aligned} \Omega_{\mu\nu\rho}^{(2)}(i, j, k) &= \\ &= -\frac{1}{8\pi^2} \varepsilon_{\mu\nu\rho\sigma} (1 + 2 \cos \alpha \cos \beta \cos \gamma - \cos^2 \alpha - \cos^2 \beta - \cos^2 \gamma)^{-1} \\ &\times \{ (\vec{\alpha} + \vec{\beta} - \vec{\gamma}) \cdot \sin \alpha \vec{e}_{\alpha} [\partial_{\sigma} (\sin \beta \vec{e}_{\beta}) \cdot \sin \gamma \vec{e}_{\gamma} - \sin \beta \vec{e}_{\beta} \cdot \partial_{\sigma} (\sin \gamma \vec{e}_{\gamma})] \\ &+ (\vec{\alpha} + \vec{\beta} - \vec{\gamma}) \cdot \sin \beta \vec{e}_{\beta} [\partial_{\sigma} (\sin \gamma \vec{e}_{\gamma}) \cdot \sin \alpha \vec{e}_{\alpha} - \sin \gamma \vec{e}_{\gamma} \cdot \partial_{\sigma} (\sin \alpha \vec{e}_{\alpha})] \\ &+ (\vec{\alpha} + \vec{\beta} - \vec{\gamma}) \cdot \sin \gamma \vec{e}_{\gamma} [\partial_{\sigma} (\sin \alpha \vec{e}_{\alpha}) \cdot \sin \beta \vec{e}_{\beta} - \sin \alpha \vec{e}_{\alpha} \cdot \partial_{\sigma} (\sin \beta \vec{e}_{\beta})] \}. \end{aligned} \quad (\text{C.4.13})$$

Here we have parametrized the transition functions as

$$v_{ij} = \exp(i\vec{\alpha} \cdot \sigma), \quad v_{jk} = \exp(i\vec{\beta} \cdot \sigma), \quad v_{ik} = \exp(i\vec{\gamma} \cdot \sigma). \quad (\text{C.4.14})$$

The cocycle condition $v_{ik} = v_{ij}v_{jk}$ then takes the form

$$\begin{aligned} \cos \gamma &= \cos \alpha \cos \beta - \sin \alpha \sin \beta \vec{e}_{\alpha} \cdot \vec{e}_{\beta}, \\ \sin \gamma \vec{e}_{\gamma} &= \sin \alpha \cos \beta \vec{e}_{\alpha} + \cos \alpha \sin \beta \vec{e}_{\beta} - \sin \alpha \sin \beta \vec{e}_{\alpha} \times \vec{e}_{\beta}. \end{aligned} \quad (\text{C.4.15})$$

These equations define a spherical triangle on the sphere S^3 with side lengths α , β and γ .

The 2-cochain also has singularities. They occur when the spherical triangle defined by the three transition functions degenerates to a great circle, i.e. when

$$(\vec{\alpha} + \vec{\beta} - \vec{\gamma}) \cdot \vec{e}_{\alpha} = (\vec{\alpha} + \vec{\beta} - \vec{\gamma}) \cdot \vec{e}_{\beta} = -(\vec{\alpha} + \vec{\beta} - \vec{\gamma}) \cdot \vec{e}_{\gamma} = 2\pi. \quad (\text{C.4.16})$$

In the neighborhood of such points the 2-cochain has a directional singularity

$$\Omega_{\mu\nu\rho}^{(2)}(i, j, k) = -\frac{1}{4\pi^2} \varepsilon_{\mu\nu\rho\sigma} (\vec{\alpha} + \vec{\beta} - \vec{\gamma}) \cdot (\vec{e}_s \times \partial_\sigma \vec{e}_s). \quad (\text{C.4.17})$$

Here we have defined another unit vector

$$\vec{e}_s = \cos \alpha \sin \beta \sin \gamma \vec{e}_\beta \times \vec{e}_\gamma + \cos \beta \sin \gamma \sin \alpha \vec{e}_\gamma \times \vec{e}_\alpha + \cos \gamma \sin \alpha \sin \beta \vec{e}_\alpha \times \vec{e}_\beta. \quad (\text{C.4.18})$$

Using Gauss' law we now obtain

$$Q_\Sigma^{(1)} = Q^{(2)} + Q_\Sigma^{(2)}, \quad (\text{C.4.19})$$

with

$$Q^{(2)} = \frac{1}{3!} \sum_{ijk} \sum_{x \in c_i \cap c_j \cap c_k} \frac{1}{4\pi^2} \int_{S_\varepsilon^1(x)} d\sigma_{\mu\nu\rho} \varepsilon_{\mu\nu\rho\sigma} (\vec{\alpha} + \vec{\beta} - \vec{\gamma}) \cdot (\vec{e}_s \times \partial_\sigma \vec{e}_s). \quad (\text{C.4.20})$$

Now $S_\varepsilon^1(x)$ is an infinitesimal circle around the singularity. Performing the integral yields

$$\frac{1}{4\pi^2} \int_{S_\varepsilon^1(x)} d\sigma_{\mu\nu\rho} \varepsilon_{\mu\nu\rho\sigma} (\vec{\alpha} + \vec{\beta} - \vec{\gamma}) \cdot (\vec{e}_s \times \partial_\sigma \vec{e}_s) = n^{(2)}(x; i, j, k), \quad (\text{C.4.21})$$

where $n^{(2)}(x; i, j, k)$ is a winding number from $\Pi_1[S^1] = \mathbb{Z}$, which counts how often the unit vector \vec{e}_s (that is perpendicular to $\vec{\alpha} + \vec{\beta} - \vec{\gamma}$) covers the circle S^1 as we integrate over $S_\varepsilon^1(x)$. Hence we obtain

$$Q^{(2)} = \sum_{\Lambda^{(2)}} n^{(2)}(x; i, j, k), \quad (\text{C.4.22})$$

where the set of singular points is now denoted by

$$\Lambda^{(2)} = \cup_{ijk} \{x \in c_i \cap c_j \cap c_k\}. \quad (\text{C.4.23})$$

The remaining contributions to the topological charge take the form

$$\begin{aligned} Q_\Sigma^{(2)} &= \frac{1}{3!} \sum_{ijk} \int_{\partial(c_i \cap c_j \cap c_k)} d\sigma_{\mu\nu\rho} \Omega_{\mu\nu\rho}^{(2)}(i, j, k) \\ &= \frac{1}{4!} \sum_{ijkl} \int_{c_i \cap c_j \cap c_k \cap c_l} d\sigma_{\mu\nu\rho} [\Omega_{\mu\nu\rho}^{(2)}(i, j, k) - \Omega_{\mu\nu\rho}^{(2)}(i, j, l) \\ &\quad + \Omega_{\mu\nu\rho}^{(2)}(i, k, l) - \Omega_{\mu\nu\rho}^{(2)}(j, k, l)]. \end{aligned} \quad (\text{C.4.24})$$

It comes as no surprise that the coboundary operator

$$\begin{aligned}\Delta\Omega_{\mu\nu\rho}^{(2)}(i, j, k, l) &= \Omega_{\mu\nu\rho}^{(2)}(i, j, k) - \Omega_{\mu\nu\rho}^{(2)}(i, j, l) + \Omega_{\mu\nu\rho}^{(2)}(i, k, l) - \Omega_{\mu\nu\rho}^{(2)}(j, k, l) \\ &= \partial_\sigma\Omega_{\mu\nu\rho\sigma}^{(3)}(i, j, k, l)\end{aligned}\quad (\text{C.4.25})$$

yields the total divergence of a 3-cochain. It is remarkable that the 3-cochain has a geometric meaning. It is given by the volume $V(i, j, k, l)$ of the spherical tetrahedron consisting of four spherical triangles defined by the participating transition functions. It is quite tedious to show that

$$\Omega_{\mu\nu\rho\sigma}^{(3)}(i, j, k, l) = \frac{1}{2\pi^2}\varepsilon_{\mu\nu\rho\sigma}V(i, j, k, l). \quad (\text{C.4.26})$$

The key observation is that the variation of the volume of the spherical tetrahedron is given by Schläfli's differential form

$$\partial_\sigma V(i, j, k, l) = \frac{1}{2}(\alpha\partial_\sigma A + \beta\partial_\sigma B + \gamma\partial_\sigma \Gamma + \delta\partial_\sigma \Delta + \varepsilon\partial_\sigma E + \zeta\partial_\sigma Z), \quad (\text{C.4.27})$$

where A, B, \dots, Z are the angles between the faces of the spherical tetrahedron defined by the transition functions. Schläfli was a mathematician in Bern (Switzerland) around the time when Einstein worked there in the patent office. We can now perform the final integration and get

$$Q_\Sigma^{(2)} = Q^{(3)} + Q_\Sigma^{(3)}. \quad (\text{C.4.28})$$

The first contribution has the form

$$Q^{(3)} = \sum_{\Lambda^{(3)}} n^{(3)}(x; i, j, k, l), \quad (\text{C.4.29})$$

where

$$\Lambda^{(3)} = \cup_{ijkl}\{x \in c_i \cap c_j \cap c_k \cap c_l\} \quad (\text{C.4.30})$$

is a set of singular points at which the surface of the spherical tetrahedron degenerates to a 2-sphere. At the singularities the volume of the spherical tetrahedron changes by an integer times $2\pi^2$ — the volume of the sphere S^3 . The local winding number

$$n^{(3)}(x; i, j, k, l) = \frac{1}{2\pi^2}\Delta V(i, j, k, l) \quad (\text{C.4.31})$$

measures the change in volume $\Delta V(i, j, k, l)$ at the singularity. The remaining

contribution takes the form

$$\begin{aligned}
Q_\Sigma^{(3)} &= \frac{1}{4!} \sum_{\partial(c_i \cap c_j \cap c_k \cap c_l)} \sigma_{\mu\nu\rho\sigma} \Omega_{\mu\nu\rho\sigma}^{(3)}(i, j, k, l) \\
&= \frac{1}{5!} \sum_{c_i \cap c_j \cap c_k \cap c_l \cap c_m} \sigma_{\mu\nu\rho\sigma} [\Omega_{\mu\nu\rho\sigma}^{(3)}(i, j, k, l) - \Omega_{\mu\nu\rho\sigma}^{(3)}(i, j, k, m) \\
&\quad + \Omega_{\mu\nu\rho\sigma}^{(3)}(i, j, l, m) - \Omega_{\mu\nu\rho\sigma}^{(3)}(i, k, l, m) + \Omega_{\mu\nu\rho\sigma}^{(3)}(j, k, l, m)] \\
&= \frac{1}{5!} \sum_{c_i \cap c_j \cap c_k \cap c_l \cap c_m} \sigma \frac{1}{2\pi^2} [V(i, j, k, l) - V(i, j, k, m) \\
&\quad + V(i, j, l, m) - V(i, k, l, m) + V(j, k, l, m)]. \tag{C.4.32}
\end{aligned}$$

Here $\sigma = \sigma_{\mu\nu\rho\sigma} \varepsilon_{\mu\nu\rho\sigma}$ determines the orientation of $\partial(c_i \cap c_j \cap c_k \cap c_l)$. The intersection of four 4-dimensional regions determines a point x . All these points form a set

$$\Lambda^{(4)} = \{x \in \cup_{ijklm} c_i \cap c_j \cap c_k \cap c_l \cap c_m\}, \tag{C.4.33}$$

such that

$$Q_\Sigma^{(3)} = Q^{(4)} = \sum_{\Lambda^{(4)}} n^{(4)}(x; i, j, k, l, m) \tag{C.4.34}$$

with

$$\begin{aligned}
n^{(4)}(x; i, j, k, l, m) &= \sigma \frac{1}{2\pi^2} [V(i, j, k, l) - V(i, j, k, m) + V(i, j, l, m) \\
&\quad - V(i, k, l, m) + V(j, k, l, m)]. \tag{C.4.35}
\end{aligned}$$

This is always an integer because the five spherical tetrahedra together are compact and cover S^3 a certain number of times — their total volume is an integer multiple of $2\pi^2$. Altogether the topological charge is a sum of local winding numbers of different topological origin

$$\begin{aligned}
Q &= \sum_{\Lambda^{(1)}} n^{(1)}(x; i, j) + \sum_{\Lambda^{(2)}} n^{(2)}(x; i, j, k) \\
&\quad + \sum_{\Lambda^{(3)}} n^{(3)}(x; i, j, k, l) + \sum_{\Lambda^{(4)}} n^{(4)}(x; i, j, k, l, m). \tag{C.4.36}
\end{aligned}$$

Again, one should note that the local winding numbers are not gauge invariant. Only their sum — the total topological charge — is gauge invariant and has a physical meaning.

C.5 The Instanton in $SU(2)$

We have argued mathematically that gauge field configurations fall into topologically distinct classes. Now we want to construct concrete examples of topologically nontrivial field configurations. Here we consider instantons, which have $Q = 1$ and are solutions of the Euclidean classical field equations. The instanton occurs at a given instant in Euclidean time. Since these solutions do not live in Minkowski space-time they have no direct interpretation in terms of real time events. Also it is unclear which role they play in the quantum theory. Instantons describe tunneling processes between degenerate classical vacuum states. Their existence gives rise to the θ -vacuum structure of non-Abelian gauge theories.

Here we concentrate on $SU(2)$. This is sufficient, because we have seen that the $SU(N)$ topological charge can be reduced to the $SU(2)$ case. In this section we go back to an infinite space with a boundary sphere S^3 , and we demand that the gauge field has finite action. Then at space-time infinity the gauge potential is in a pure gauge

$$G_\mu(x) = g(x)\partial_\mu g(x)^\dagger. \quad (\text{C.5.1})$$

Provided the gauge field is otherwise smooth, the topology resides entirely in the mapping g . We want to construct a field configuration with topological charge $Q = 1$, i.e. one in which the mapping g covers the group space $SU(2) = S^3$ once as we integrate over the boundary sphere S^3 . The simplest mapping of this sort is the identity, i.e. each point at the boundary of space-time is mapped into the corresponding point in the group space such that

$$g(x) = \frac{x_0 + i\vec{x} \cdot \vec{\sigma}}{|x|}, \quad |x| = \sqrt{x_0^2 + |\vec{x}|^2}. \quad (\text{C.5.2})$$

Next we want to extend the gauge field to the interior of space-time without introducing singularities. We cannot simply maintain the form of eq.(C.5.1) because g is singular at $x = 0$. To avoid this singularity we make the ansatz

$$G_\mu(x) = f(|x|)g(x)\partial_\mu g(x)^\dagger, \quad (\text{C.5.3})$$

where $f(\infty) = 1$ and $f(0) = 0$. For any smooth function f with these properties the above gluon field configuration has $Q = 1$. Still, this does not mean that we have constructed an instanton. Instantons are field configurations with $Q \neq 0$ that are in addition solutions of the Euclidean classical equations of motion, i.e. they are minima of the Euclidean action

$$S[G_\mu] = \int d^4x \frac{1}{2g^2} \text{Tr}[G_{\mu\nu}(x)G_{\mu\nu}(x)]. \quad (\text{C.5.4})$$

Let us consider the following integral

$$\begin{aligned} \int d^4x \operatorname{Tr}[(G_{\mu\nu}(x) \pm \frac{1}{2}\varepsilon_{\mu\nu\rho\sigma}G_{\rho\sigma}(x))(G_{\mu\nu}(x) \pm \frac{1}{2}\varepsilon_{\mu\nu\kappa\lambda}G_{\kappa\lambda}(x))] = \\ \int d^4x \operatorname{Tr}[G_{\mu\nu}(x)G_{\mu\nu}(x) \pm \varepsilon_{\mu\nu\rho\sigma}G_{\mu\nu}(x)G_{\rho\sigma}(x) + G_{\mu\nu}(x)G_{\mu\nu}(x)] = \\ 4g_s^2 S[G_\mu] \pm 32\pi^2 Q[G_\mu]. \end{aligned} \quad (\text{C.5.5})$$

We have integrated a square. Hence it is obvious that

$$S[G_\mu] \pm \frac{8\pi^2}{g^2} Q[G_\mu] \geq 0 \Rightarrow S[G_\mu] \geq \frac{8\pi^2}{g^2} |Q[G_\mu]|, \quad (\text{C.5.6})$$

i.e. a topologically nontrivial field configuration costs at least a minimum action proportional to the topological charge. Instantons are configurations with minimum action, i.e. for them

$$S[G_\mu] = \frac{8\pi^2}{g^2} |Q[G_\mu]|. \quad (\text{C.5.7})$$

From the above argument it is clear that a minimum action configuration arises only if

$$G_{\mu\nu}(x) = \pm \frac{1}{2}\varepsilon_{\mu\nu\rho\sigma}G_{\rho\sigma}(x). \quad (\text{C.5.8})$$

Configurations that obey this equation with a plus sign are called selfdual. The ones that obey it with a minus sign are called anti-selfdual. It is instructive to convince oneself that the above gluon field with

$$f(|x|) = \frac{|x|^2}{|x|^2 + \rho^2} \quad (\text{C.5.9})$$

is indeed an instanton for any value of ρ . The instanton configuration hence takes the form

$$G_\mu(x) = \frac{|x|^2}{|x|^2 + \rho^2} g(x) \partial_\mu g(x)^\dagger. \quad (\text{C.5.10})$$

There is a whole family of instantons with different radii ρ . As a consequence of scale invariance of the classical action they all have the same action $S[G_\mu] = 8\pi^2/g^2$.

C.6 θ -Vacua

The existence of topologically nontrivial gauge transformations has drastic consequences for non-Abelian gauge theories. In fact, there is not just one classical

vacuum state, but there is one for each topological winding number. Instantons describe tunneling transitions between topologically distinct vacua. Due to tunneling the degeneracy of the classical vacuum states is lifted, and the true quantum vacuum turns out to be a θ -state, i.e. one in which configurations of different winding numbers are mixed.

In the following we fix to $G_4(x) = 0$ gauge, and we consider space to be compactified from \mathbb{R}^3 to S^3 . This is just a technical trick which makes life easier. Using transition functions one could choose any other compactification, e.g. on a torus T^3 , or one could choose appropriate boundary conditions on \mathbb{R}^3 itself. The classical vacuum solutions of such a theory are the pure gauge fields

$$G_i(x) = g(x)\partial_i g(x)^\dagger. \quad (\text{C.6.1})$$

Since we have compactified space the classical vacua can be classified by their winding number

$$n \in \Pi_3[SU(3)] = \mathbb{Z}, \quad (\text{C.6.2})$$

which is given by

$$n = \frac{1}{24\pi^2} \int_{S^3} d^3x \, \varepsilon_{ijk} \text{Tr}[g(x)\partial_i g(x)^\dagger g(x)\partial_j g(x)^\dagger g(x)\partial_k g(x)^\dagger]. \quad (\text{C.6.3})$$

One might think that one can construct a quantum vacuum $|n\rangle$ just by considering small fluctuations around a classical vacuum with given n . Quantum tunneling, however, induces transitions between the various classical vacua. Imagine the system is in a classical vacuum state with winding number m at early times $t = -\infty$, then it changes continuously (now deviating from a pure gauge), and finally at $t = \infty$ it returns to a classical vacuum state with a possibly different winding number n . The time evolution corresponds to one particular path in the Feynman path integral. The corresponding gauge field smoothly interpolates between the initial and final classical vacua. When we calculate its topological charge, we can use Gauss' law, which yields an integral of the 0-cochain over the space-time boundary, which consists of the spheres S^3 at $t = -\infty$ and at $t = \infty$. At each boundary sphere the gauge field is in a pure gauge, and the integral yields the corresponding winding number such that

$$Q = n - m. \quad (\text{C.6.4})$$

Hence, a configuration with topological charge Q induces a transition from a classical vacuum with winding number m to one with winding number $n = m + Q$. In other words, the Feynman path integral that describes the amplitude for transitions from one classical vacuum to another is restricted to field configurations

in the topological sector Q , such that

$$\langle n|U(\infty, -\infty)|m\rangle = \int \mathcal{D}G_\mu^{(n-m)} \exp(-S[G_\mu]). \quad (\text{C.6.5})$$

Here $G_\mu^{(Q)}$ denotes a gauge field with topological charge Q , and $U(t', t)$ is the time evolution operator.

It is crucial to note that the winding number n is not gauge invariant. In fact, as we perform a gauge transformation with winding number 1 the winding number of the pure gauge field changes to $n + 1$. In the quantum theory such a gauge transformation g is implemented by a unitary operator T that acts on wave functionals $\Psi[G_i]$ by gauge transforming the field G_i , i.e.

$$T\Psi[G_i] = \Psi[g(G_i + \partial_i)g^\dagger]. \quad (\text{C.6.6})$$

In particular, acting on a state that describes small fluctuations around a classical vacuum one finds

$$T|n\rangle = |n + 1\rangle, \quad (\text{C.6.7})$$

i.e. T acts as a ladder operator. Since the operator T implements a special gauge transformation, it commutes with the Hamiltonian, just the theory is gauge invariant. This means that the Hamiltonian and T can be diagonalized simultaneously, and each eigenstate can be labelled by an eigenvalue of T . Since T is a unitary operator its eigenvalues are complex phases $\exp(i\theta)$, such that an eigenstate — for example the vacuum — can be written as $|\theta\rangle$ with

$$T|\theta\rangle = \exp(i\theta)|\theta\rangle. \quad (\text{C.6.8})$$

On the other hand, we can construct the θ -vacuum as a linear combination

$$|\theta\rangle = \sum_n c_n |n\rangle. \quad (\text{C.6.9})$$

Using

$$\begin{aligned} T|\theta\rangle &= \sum_n c_n T|n\rangle = \sum_n c_n |n + 1\rangle \\ &= \sum_n c_{n-1} |n\rangle = \exp(i\theta) \sum_n c_n |n\rangle, \end{aligned} \quad (\text{C.6.10})$$

one obtains $c_{n-1} = \exp(i\theta)c_n$ such that $c_n = \exp(-in\theta)$ and

$$|\theta\rangle = \sum_n \exp(-in\theta) |n\rangle. \quad (\text{C.6.11})$$

The true vacuum of a non-Abelian gauge theory is a linear combination of classical vacuum states of different winding numbers. For each value of θ there is a corresponding vacuum state. This is analogous to the energy bands in a solid. There a state is labelled by a Bloch momentum as a consequence of the discrete translation symmetry. In non-Abelian gauge theories T induces discrete translations between classical vacua, with analogous mathematical consequences.

Now let us consider the quantum transition amplitude between different θ -vacua

$$\begin{aligned}
\langle \theta | U(\infty, -\infty) | \theta' \rangle &= \sum_{m,n} \exp(in\theta) \exp(-im\theta') \langle n | U(\infty, -\infty) | m \rangle \\
&= \sum_{n, Q=n-m} \exp(in\theta - i(n-Q)\theta') \int \mathcal{D}G_\mu^{(Q)} \exp(-S[G_\mu]) \\
&= \delta(\theta - \theta') \sum_Q \int \mathcal{D}G_\mu^{(Q)} \exp(-S[G_\mu]) \exp(i\theta Q[G_\mu]) \\
&= \int \mathcal{D}G_\mu \exp(-S_\theta[G_\mu]).
\end{aligned} \tag{C.6.12}$$

There is no transition between different θ -vacua, which confirms that they are eigenstates. Also we can again identify the action in a θ -vacuum as

$$S_\theta[G_\mu] = S[G_\mu] - i\theta Q[G_\mu]. \tag{C.6.13}$$

Finally, let us consider the theory with at least one massless fermion. In that case the Dirac operator $\gamma_\mu(G_\mu(x) + \partial_\mu)$ has a zero mode. This follows from an index theorem due to Atiyah and Singer. They considered the eigenvectors of the Dirac operator with zero eigenvalue

$$\gamma_\mu(G_\mu(x) + \partial_\mu)\Psi(x) = 0. \tag{C.6.14}$$

These eigenvectors have a definite handedness, i.e.

$$\frac{1}{2}(1 \pm \gamma_5)\Psi(x) = \Psi(x), \tag{C.6.15}$$

because

$$\gamma_5 \gamma_\mu(G_\mu(x) + \partial_\mu)\Psi(x) = -\gamma_\mu(G_\mu(x) + \partial_\mu)\gamma_5\Psi(x) = 0. \tag{C.6.16}$$

The Atiyah-Singer index theorem states that

$$Q = n_L - n_R, \tag{C.6.17}$$

where n_L and n_R are the numbers of left and right handed zero modes. Hence, a topologically nontrivial gauge field configuration necessarily has at least one zero mode. This zero mode of the Dirac operator eliminates topologically nontrivial field configurations from theories with massless fermions, i.e. then $Q[G_\mu] = 0$ for all configurations that contribute to the Feynman path integral. In that case the θ -term in the action has no effect, and all θ -vacua would be physically equivalent. This scenario has been suggested as a possible solution of the strong CP problem. If the lightest quark (the u quark) would be massless, θ would not generate an electric dipole moment for the neutron. There is still no agreement on this issue. Some experts of chiral perturbation theory claim that a massless u-quark is excluded by experimental data. However, the situation is not clear. For example, the pion mass depends only on the sum $m_u + m_d$, and one must look at more subtle effects. My personal opinion is that the solution of the strong CP problem is beyond QCD, and probably beyond the Standard model, and is maybe related to some cosmological effect.